Mathematics of Robustness, Risk & Failure

Volume I: Extreme Values & Tail Risk

Heavy tails, power laws, and the mathematics of rare but consequential events

---

Every derivation step by step

Every symbol defined

Explicit numerical examples

Geometric intuitions

Silvano Salvador

Contents

1.1 Order Statistics and the Behaviour of Maxima 5

1.2 The Three Extreme Value Domains: Gumbel, Fréchet, Weibull 8

1.3 The Generalized Extreme Value (GEV) Distribution 11

1.4 Domains of Attraction and the Max-Stable Property 14

1.5 Block Maxima Method: Estimation and Diagnostics 17

2.1 The Generalized Pareto Distribution (GPD) 21

2.2 The Pickands–Balkema–de Haan Theorem 23

2.3 Threshold Selection: Mean Excess Plots and Hill Estimator 26

2.4 Maximum Likelihood and Probability-Weighted Moment Estimation 30

2.5 Return Levels and Return Periods 33

3.1 Regularly Varying Functions and Karamata's Theorem 36

3.2 Power-Law Distributions: Definition, Estimation, and Pitfalls 39

3.3 Subexponential Distributions and the Principle of a Single Big Jump 42

3.4 Stable Distributions and Generalized Central Limit Theorems 45

3.5 Multivariate Extremes and Tail Dependence 48

4.1 Value-at-Risk (VaR) and Its Limitations 51

4.2 Conditional Value-at-Risk (CVaR) / Expected Shortfall 54

4.3 Coherent and Convex Risk Measures: Axiomatic Foundations 57

4.4 Spectral and Distortion Risk Measures 61

4.5 Backtesting and Elicitability of Risk Measures 64

A.1 The Delta Method for Extreme Value Functionals 68

A.2 The r-largest Order Statistics Method 69

A.3 Worked Exercises for Part 1 70

B.1 Declustering and the Extremal Index 73

B.2 The Poisson–GPD Model in Full 74

B.3 Worked Exercises for Part 2 75

C.1 Second-Order Regular Variation and Bias-Corrected Estimators 77

C.2 Extremes of Dependent Sequences: Beyond i.i.d. 78

C.3 Worked Exercises for Part 3 79

D.1 Estimating Risk Measures Under Parameter Uncertainty: A Bayesian Treatment 81

D.2 Risk Measure Aggregation and the Subadditivity–Diversification Distinction 82

D.3 Worked Exercises for Part 4 83

D.4 Synthesis: Extreme Value Theory as the Foundation for Volume I 84

E.1 Case Study: Design Flood Estimation for a River Levee 86

E.2 Case Study: Catastrophe Reinsurance Pricing 87

E.3 Case Study: Market Risk Capital Under Basel FRTB 89

E.4 Case Study: Fatigue Life Extreme Quantiles for a Critical Component 90

E.5 Case Study: Seismic Hazard Assessment — the Gutenberg–Richter Law as a Power Law 92

F.1 Full Proof: The Cauchy Functional Equation in the FTG Theorem 95

F.2 Full Derivation: GEV Moments via the Gamma Function 96

F.3 Comprehensive Formula Reference 97

F.4 Glossary of Key Terms 98

G.1 Problem Setup 100

G.2 Severity Modelling 100

G.3 Frequency Modelling 101

G.4 Aggregate Loss Distribution via Monte Carlo 101

G.5 Uncertainty Quantification: Putting It All Together 102

H.1 Historical Development of Extreme Value Theory 104

H.2 Historical Development of Coherent Risk Measures 104

H.3 Connections to Adjacent Fields Not Covered in This Volume 105

H.4 Suggested Further Reading 106

I.1 Covariate-Dependent GEV and GPD Models 108

I.2 Three Competing Notions of Return Period Under Non-Stationarity 109

I.3 Non-Stationary Risk Measures and Their Aggregation 111

J.1 Parametric Copula Families and Their Tail Behaviour 113

J.2 Estimation: Maximum Likelihood, IFM, and Rank-Based Methods 114
J.3 Goodness-of-Fit Testing for Copulas 115
J.4 Integrated Worked Example: Two-Asset Portfolio Tail Risk 116
K.1 Prior Elicitation for the Shape Parameter 119
K.2 Metropolis–Hastings for the GPD Posterior 120
K.3 Posterior Predictive Risk Measures 122
L.1 Probability Spaces and Measurability 124
L.2 Convergence Concepts 124
L.3 Characteristic Functions and the Continuity Theorem 125
L.4 Key Results from Real Analysis 126
L.5 Regular Variation: Technical Supplement 126
Extreme Value Theory: Foundations 128
Regular Variation, Heavy Tails, and Subexponential Distributions 128
Coherent Risk Measures and Financial Risk 129
Copulas and Dependence Modelling 129
Threshold Methods, Estimation, and Applications 130
Bayesian Methods in Extreme Value Analysis 131
Non-Stationarity, Climate Extremes, and Operational Risk 131
Robust Optimization and Decision under Uncertainty 131
L.1 Probability Spaces and Measurability 133
L.2 Convergence Concepts 133
L.3 Characteristic Functions and the Continuity Theorem 134
L.4 Key Results from Real Analysis 135
L.5 Regular Variation: Technical Supplement 135
Extreme Value Theory: Foundations 137
Regular Variation, Heavy Tails, and Subexponential Distributions 137
Coherent Risk Measures and Financial Risk 138
Copulas and Dependence Modelling 138

Threshold Methods, Estimation, and Applications **139**

Bayesian Methods in Extreme Value Analysis **140**

Non-Stationarity, Climate Extremes, and Operational Risk **140**

Robust Optimization and Decision under Uncertainty **140**

PART 1

Foundations of Extreme Value Theory

We build the classical theory of maxima and tails: the Fisher–Tippett–Gnedenko theorem, domains of attraction, and the generalized extreme value (GEV) distribution.

1.1 Order Statistics and the Behaviour of Maxima

Overview

This section establishes the probabilistic machinery for studying the extreme values of a sample: order statistics, their exact and asymptotic distributions, and the first hints of why the *maximum* of n independent observations behaves so differently from their *mean*. Everything that follows in this volume — the Fisher–Tippett–Gnedenko theorem, the generalized Pareto distribution, power laws, and coherent risk measures — is ultimately a refinement of the questions raised here.

Order statistics. Let $X_1, \dots, X_n$ be i.i.d. random variables with common cumulative distribution function (CDF) F and, where it exists, density f. Arrange them in increasing order:

$$X_{(1)} \le X_{(2)} \le \dots \le X_{(n)}.$$

The random variable $X_{(k)}$ is the *k-th order statistic*. Two cases dominate this volume: the *sample maximum* $M_n = X_{(n)} = \max(X_1, \dots, X_n)$ and, dually, the sample minimum $X_{(1)}$ (which is recovered from the maximum by the substitution $X \to -X$, so all results for maxima transfer immediately to minima).

Definition (Exact distribution of the maximum)

For i.i.d. $X_1, \dots, X_n \sim F$, the CDF of $M_n = \max(X_1, \dots, X_n)$ is

$$F_{M_n}(x) = P(M_n \le x) = P(X_1 \le x, \dots, X_n \le x) = \prod_{i=1}^{n} P(X_i \le x) = F(x)^n.$$

This identity — $F_{M_n}(x) = F(x)^n$ — is deceptively simple, yet it is the entire foundation of extreme value theory. Everything else in this Part is devoted to understanding the *limiting behaviour* of $F(x)^n$ as $n \to \infty$ after an appropriate change of location and scale, because $F(x)^n$ itself degenerates: for any fixed x with $F(x) < 1$, $F(x)^n \to 0$, and for $F(x) = 1$, $F(x)^n = 1$. The maximum, unrescaled, converges in probability to the right endpoint of the support,

$$x^* = \sup\{x : F(x) < 1\} \in (-\infty, +\infty],$$

which may be finite (e.g. Uniform$(0,1)$, $x^* = 1$) or infinite (e.g. Exponential, Normal, Pareto, $x^* = +\infty$).

Density of order statistics. If F has density f, the joint density of all n order statistics is

$$f_{X_{(1)},\dots,X_{(n)}}(x_1, \dots, x_n) = n! \prod_{i=1}^{n} f(x_i), \qquad x_1 < x_2 < \dots < x_n,$$

the factor $n!$ counting the number of permutations of the original sample consistent with a given ordering. Marginalizing, the density of the k-th order statistic is

$$f_{X_{(k)}}(x) = \frac{n!}{(k-1)!\,(n-k)!}\, F(x)^{k-1}\,[1 - F(x)]^{n-k}\, f(x).$$

For $k = n$ this reduces to $f_{M_n}(x) = n\,F(x)^{n-1} f(x)$, the derivative of $F(x)^n$, as expected.

Example — Maximum of Uniform random variables

Let $X_1, \dots, X_n \sim \text{Uniform}(0,1)$, so $F(x) = x$ on $[0,1]$. Then

$$F_{M_n}(x) = x^n, \qquad f_{M_n}(x) = nx^{n-1}, \qquad x \in [0,1].$$

The mean of M_n is

$$\mathbb{E}[M_n] = \int_0^1 x \cdot nx^{n-1}\,dx = \frac{n}{n+1} \to 1 = x^*.$$

More informatively, consider the rescaled variable $Y_n = n(1 - M_n)$, measuring the "distance to the boundary" on a scale of order $1/n$. Its CDF is

$$P(Y_n \le y) = P(M_n \ge 1 - \tfrac{y}{n}) = 1 - \Big(1 - \frac{y}{n}\Big)^n \longrightarrow 1 - e^{-y}, \qquad y \ge 0,$$

using $(1 - y/n)^n \to e^{-y}$. Thus $n(1 - M_n) \xrightarrow{d} \text{Exponential}(1)$. This is our first encounter with a *non-degenerate limit law for an extreme*, obtained by linear rescaling $a_n(M_n - b_n)$ with $a_n = n$, $b_n = 1$. The limiting distribution here is (the reflection of) a *Weibull-type* extreme value law, one of the three universal families introduced in Section 1.2.

Why the mean and the maximum diverge. For the sample mean, the Central Limit Theorem (CLT) guarantees, under finite variance, that $\sqrt{n}(\bar{X}_n - \mu) \xrightarrow{d} \mathcal{N}(0, \sigma^2)$ — a *single* universal limit (the Gaussian), governed only by σ^2, regardless of the finer shape of F. The maximum obeys an analogous but richer principle: there exist normalizing sequences $a_n > 0$, $b_n \in \mathbb{R}$ such that

$$\frac{M_n - b_n}{a_n} \xrightarrow{d} G$$

for some non-degenerate G, but *which* G depends on the tail behaviour of F in a way that the CLT does not depend on the bulk of F. The classification of all possible limits G — there are exactly three types up to location/scale — is the content of the Fisher–Tippett–Gnedenko theorem (Section 1.2).

> **A recurring intuition.** The sample mean is governed by the *typical* behaviour of X — by the bulk of the distribution, summarized by μ and σ^2. The sample maximum is governed entirely by the *tail* behaviour of X — by how $F(x) \to 1$ as $x \to x^*$. Two distributions with identical means and variances can have radically different extreme-value behaviour if their tails differ. This is the central reason a separate mathematical theory — rather than a corollary of the CLT — is required for risk, reliability, and failure analysis.

Asymptotic independence of extremes and the bulk. A classical and useful fact, due essentially to the structure of order statistics, is that for large n, the upper order statistics $X_{(n)}, X_{(n-1)}, \dots, X_{(n-k+1)}$ (the k largest values, for fixed k) become, after appropriate normalization, asymptotically independent of the bulk of the sample. This underlies the validity of *block maxima* and *peaks-over-threshold* approaches (Section 1.5 and Chapter 2): one may model the extreme tail separately from the rest of the data without contaminating either analysis, provided the asymptotic regime is approached carefully.

Return to the general convergence problem. We seek conditions on F under which con-

stants $a_n > 0$, b_n exist with

$$F(a_n x + b_n)^n \longrightarrow G(x) \qquad \text{for all } x \text{ at which } G \text{ is continuous,}$$

for some non-degenerate CDF G. Equivalently, writing $u_n(x) = a_n x + b_n$ and using $1 - p \approx -\log p$ for p near 1,

$$n\,[1 - F(u_n(x))] \longrightarrow -\log G(x),$$

so the convergence of F^n is governed by the rate at which the *survival function* $\overline{F} = 1 - F$ decays near x^*. The classification of these decay rates — exponential-type (Gumbel domain), power-type (Fréchet domain), or bounded-support with polynomial approach (Weibull domain) — is precisely the subject of the next section.

A second derivation via the exponential approximation. There is an illuminating alternative route to the limit $n(1 - M_n) \xrightarrow{d} \text{Exp}(1)$ for the Uniform case, one that generalizes far more readily than the direct CDF computation given above. Write $X_{(n)} = 1 - \frac{Y_n}{n}$ for some random variable $Y_n \geq 0$. For $X_i \sim \text{Uniform}(0, 1)$, the event $\{X_i > 1 - y/n\}$ has probability y/n. The number of the n points $X_1, \ldots, X_n$ falling in the interval $(1 - y/n, 1]$ is therefore $\text{Binomial}(n, y/n)$, which converges in distribution to $\text{Poisson}(y)$ as $n \to \infty$ (the classical Poisson limit theorem for rare events). The event $\{Y_n > y\} = \{n(1 - M_n) > y\} = \{M_n < 1 - y/n\}$ is exactly the event that *none* of the n points fall in $(1 - y/n, 1]$, i.e.

$$P(Y_n > y) \to P(\text{Poisson}(y) = 0) = e^{-y},$$

recovering $Y_n \xrightarrow{d} \text{Exp}(1)$. **This Poisson-points argument is the conceptual seed of the entire Poisson-process formulation of extreme value theory** developed in Appendix B.2: the limiting behaviour of extremes is governed by the asymptotic Poisson distribution of the *count* of points exceeding a high threshold, a perspective that generalizes immediately to non-uniform F (via the probability integral transform $U_i = F(X_i) \sim \text{Uniform}(0, 1)$) and to the peaks-over-threshold framework of Chapter 2, where the "threshold" plays the role of $1 - y/n$ above.

Example — The same argument for a general $F \in D(G_\xi)$

For general F, suppose a_n, b_n are normalizing constants such that $n[1 - F(a_n x + b_n)] \to \tau(x)$ for some function τ (this is precisely the condition underlying the FTG theorem, Section 1.2). The number of points X_i exceeding the level $a_n x + b_n$ is, by the same Poisson-limit argument, asymptotically $\text{Poisson}(\tau(x))$, and

$$P(M_n \leq a_n x + b_n) = P(\text{no exceedances}) \to e^{-\tau(x)}.$$

For this to be a valid (non-degenerate) limit $G(x) = e^{-\tau(x)}$, the function $\tau(x)$ must be such that $e^{-\tau(x)}$ is a CDF — τ decreasing from $+\infty$ to 0. The three forms of $\tau(x)$ consistent with $\tau(\lambda x) = \lambda^{-\theta}\tau(x)$-type scaling (homogeneity, the discrete analogue of the max-stability functional equation of Section 1.2's proof) correspond precisely to $\tau(x) = e^{-x}$ (Gumbel), $\tau(x) = x^{-\alpha}$ (Fréchet), and $\tau(x) = (-x)^{\alpha}$ (Weibull) — the three GEV types, now derived via a probabilistic (Poisson) argument rather than the purely functional-equation approach of Section 1.2's proof. This dual derivation — analytic (functional equations) versus probabilistic (Poisson approximation) — recurs throughout extreme value theory, and the probabilistic route is often more readily generalized to the multivariate and process-level settings of Section 3.5 and Appendix C.2.

1.2 The Three Extreme Value Domains: Gumbel, Fréchet, Weibull

Overview

We state and prove the Fisher–Tippett–Gnedenko (FTG) theorem, the central limit theorem of extreme value theory: the only possible non-degenerate limits of normalized maxima are three families (Gumbel, Fréchet, Weibull), unifiable into the Generalized Extreme Value distribution. We characterize, via the tail of F, which family applies to common distributions.

The max-stability property. A non-degenerate CDF G is called *max-stable* if, for every $n \geq 1$, there exist constants $\alpha_n > 0$, $\beta_n \in \mathbb{R}$ such that

$$G(x)^n = G(\alpha_n x + \beta_n) \qquad \text{for all } x.$$

Max-stability is the defining property of any possible limit law for normalized maxima: if M_n (suitably normalized) converges to G, then M_{kn} for any k must, after its own normalization, converge to the same G — and the relationship between the normalizations of M_{kn} and M_n forces precisely the functional equation above.

Theorem (Fisher–Tippett–Gnedenko, 1928/1943)

If there exist sequences $a_n > 0$, $b_n \in \mathbb{R}$ and a non-degenerate CDF G such that

$$P\left(\frac{M_n - b_n}{a_n} \leq x\right) = F(a_n x + b_n)^n \longrightarrow G(x) \quad \text{as } n \to \infty,$$

then G belongs (up to location-scale transformations $x \mapsto (x-\mu)/\sigma$) to exactly one of the following three families:

$$\textbf{Gumbel:} \quad \Lambda(x) = \exp(-e^{-x}), \quad x \in \mathbb{R};$$

$$\textbf{Fréchet:} \quad \Phi_\alpha(x) = \begin{cases} 0, & x \leq 0, \\ \exp(-x^{-\alpha}), & x > 0, \end{cases} \qquad \alpha > 0;$$

$$\textbf{Weibull:} \quad \Psi_\alpha(x) = \begin{cases} \exp(-(-x)^{\alpha}), & x \leq 0, \\ 1, & x > 0, \end{cases} \qquad \alpha > 0.$$

Proof sketch. The proof proceeds in three stages.

(1) *G must be max-stable.* Fix $k \in \mathbb{N}$. Since $F^{kn}(a_{kn}x + b_{kn}) \to G(x)$ and also $F^{kn}(a_{kn}x + b_{kn}) = [F^n(a_{kn}x + b_{kn})]^k$, and separately $F^n(a_n x + b_n) \to G(x)$ implies (by a standard convergence-of-types argument) that $F^n(a_{kn}x + b_{kn}) \to G((x - \beta_k)/\alpha_k)$ for some constants $\alpha_k > 0, \beta_k$ depending on k. Combining,

$$G(x)^{1/k} \text{ “=” } G\left(\frac{x - \beta_k}{\alpha_k}\right)^{?}$$

After careful bookkeeping (we omit the convergence-of-types lemma details, which guarantee that normalizing constants are essentially unique up to asymptotic equivalence), one arrives at

$$G(x)^k = G(\alpha_k x + \beta_k)$$

for all $k \in \mathbb{N}$, and by a density/continuity argument this extends to all real $t > 0$:

$$G(x)^t = G(\alpha(t)x + \beta(t)), \qquad t > 0.$$

(2) *Solving the functional equation.* The functions $\alpha(t) > 0$ and $\beta(t)$ must satisfy the cocycle conditions $\alpha(ts) = \alpha(t)\alpha(s)$ and $\beta(ts) = \alpha(t)\beta(s) + \beta(t)$, which are Cauchy-type functional equations. Under the mild regularity inherited from G being a CDF (monotonicity), the only solutions are

$$\alpha(t) = t^{-\theta} \text{ for some } \theta \in \mathbb{R}, \qquad \beta(t) = c\,\frac{1-\alpha(t)}{\theta} \text{ (or } \beta(t) = c\log t \text{ if } \theta = 0).$$

(3) *Three cases by sign of θ.* Substituting back into $G(x)^t = G(\alpha(t)x + \beta(t))$ and solving the resulting ODE for $\log(-\log G(x))$ (after the appropriate location-scale normalization) yields:

- $\theta = 0$: $G(x) = \exp(-e^{-x})$ — the **Gumbel** law, with support $\mathbb{R}$.
- $\theta > 0$ (write $\theta = 1/\alpha$): $G(x) = \exp(-x^{-\alpha})$ for $x > 0$ — the **Fréchet** law, with support $(0, \infty)$ after normalization.
- $\theta < 0$ (write $\theta = -1/\alpha$): $G(x) = \exp(-(-x)^{\alpha})$ for $x < 0$ — the **Weibull** law, with support $(-\infty, 0)$.

This exhausts all max-stable non-degenerate distributions. ■

Heuristic characterization by tail type. The classification has a direct, checkable correspondence with the tail of F near $x^* = \sup\{x : F(x) < 1\}$:

Domain	Tail behaviour of $1 - F(x)$ as $x \to x^*$	Examples
Gumbel ($\theta = 0$)	"Exponential-type": $1 - F(x) \sim c\,e^{-x/\sigma}$ (light tails), x^* may be $\pm\infty$	Normal, Exponer
Fréchet ($\theta > 0$)	"Power-type": $1 - F(x) \sim c\,x^{-\alpha}$ as $x \to +\infty$ (heavy tails), $x^* = +\infty$	Pareto, Cauchy, S
Weibull ($\theta < 0$)	$x^* < \infty$ finite and $1 - F(x) \sim c\,(x^* - x)^{\beta}$ (bounded support)	Uniform, Beta

Example — Identifying domains for standard distributions

(a) Exponential(λ). $\overline{F}(x) = e^{-\lambda x}$. With $a_n = 1/\lambda$, $b_n = \frac{\ln n}{\lambda}$,

$$F(a_n x + b_n)^n = \left(1 - \frac{e^{-\lambda x}}{n}\right)^n \to e^{-e^{-x}} = \Lambda(x).$$

Exponential is in the **Gumbel** domain — as is every distribution with an exponentially decaying tail.

(b) Pareto, $\overline{F}(x) = x^{-\alpha}$ for $x \geq 1$. With $a_n = n^{1/\alpha}$, $b_n = 0$,

$$F(a_n x)^n = \left(1 - \frac{x^{-\alpha}}{n}\right)^n \to e^{-x^{-\alpha}} = \Phi_\alpha(x), \quad x > 0.$$

Pareto is in the **Fréchet** domain with shape α. Note that α here is exactly the power-law tail index, foreshadowing Chapter 3.

(c) Uniform$(0,1)$. As computed in Section 1.1, $n(1 - M_n) \xrightarrow{d} \text{Exp}(1)$, i.e. M_n rescaled converges to $1 - \text{Exp}(1)$, which is (up to sign and shift) $\Psi_1(x) = e^{-(-x)}$ for $x \leq 0$ — the **Weibull** domain with shape $\alpha = 1$.

A common misconception. "Weibull" here refers to the *extreme value* Weibull distribution (the limit law for *minima* of bounded-below variables, or maxima of bounded-

above variables), which is distinct from — though related by a logarithmic transform to — the *Weibull distribution* used as a lifetime model in reliability theory (Chapter 5–6). The reliability-Weibull is, in fact, in the Gumbel domain for maxima of i.i.d. lifetimes; the connection is that $-\log X$ for $X \sim$Weibull(reliability) follows a Gumbel-type law. We will be careful to disambiguate these usages when both appear, particularly in Volume II.

Significance for risk and failure analysis. The Fréchet domain is of paramount practical importance: it is the *only* domain in which the limiting distribution itself has a heavy (power-law) tail, meaning extreme outcomes remain plausible even in the limit, with moments above order α failing to exist for the limit law itself. Financial returns, insurance losses, earthquake magnitudes, and many network-traffic and biological quantities empirically fall in or near the Fréchet domain — a fact with profound consequences for risk measurement (Chapter 4) precisely because the "rare event" does not become negligible as the sample size grows; it merely becomes rarer in frequency but not in eventual relative magnitude.

The convergence-of-types theorem, stated precisely. The proof sketch of Section 1.2 invoked, without full statement, a result asserting that normalizing constants are "essentially unique." We now state this precisely, since it is used repeatedly (and silently) throughout extreme value theory whenever different parametrizations or normalizations are compared.

Theorem (Convergence of types)

Suppose $F_n(a_n x + b_n) \to G(x)$ and $F_n(\alpha_n x + \beta_n) \to G^*(x)$, where G, G^* are both non-degenerate CDFs, $a_n, \alpha_n > 0$, and $b_n, \beta_n \in \mathbb{R}$. Then $a_n/\alpha_n \to A \in (0, \infty)$ and $(b_n - \beta_n)/\alpha_n \to B \in \mathbb{R}$, and $G^*(x) = G(Ax + B)$ for all x.

In words: *if two sequences of normalizing constants both produce non-degenerate limits for the same sequence F_n, those limits differ only by a location-scale transformation, and the normalizing constants themselves are asymptotically proportional (up to the same location-scale relationship).* This is what permits statements like "the normalizing constants are unique up to asymptotically negligible terms" — a fact used implicitly whenever, e.g., one writes "$a_n = n^{1/\alpha}$" for the Fréchet domain (Section 1.2's worked example) without specifying an exact constant of proportionality, since any constant multiple ca_n would work equally well, merely rescaling the limiting Fréchet distribution's scale parameter.

Example — Illustrating non-uniqueness and its resolution

For Pareto(α, $x_{\min} = 1$), Section 1.2 used $a_n = n^{1/\alpha}$, $b_n = 0$, obtaining $F^n(a_n x) \to \Phi_\alpha(x) = e^{-x^{-\alpha}}$. Suppose instead one uses $a'_n = 2n^{1/\alpha}$ (twice as large). Then

$$F^n(a'_n x) = F^n(2n^{1/\alpha} x) = \left(1 - \frac{(2n^{1/\alpha}x)^{-\alpha}}{n}\right)^n = \left(1 - \frac{2^{-\alpha}x^{-\alpha}}{n}\right)^n \to e^{-2^{-\alpha}x^{-\alpha}} = \Phi_\alpha(2^{-1}x)?$$

More carefully, $e^{-2^{-\alpha}x^{-\alpha}} = e^{-(2x)^{-\alpha}\cdot 2^{\alpha}\cdot 2^{-\alpha}}$ — let us just verify directly: we want this to equal $\Phi_\alpha(Ax + B)$ for some A, B. Since $e^{-2^{-\alpha}x^{-\alpha}} = e^{-(2^{1/\alpha}\cdot)^{-\alpha}x^{-\alpha}}$... setting $y = 2^{-1/\alpha}$. hmm — directly: $2^{-\alpha}x^{-\alpha} = (2x)^{-\alpha}\cdot 2^{\alpha}\cdot 2^{-\alpha}$, this is circular; instead note $2^{-\alpha}x^{-\alpha} = (x/2^{?})$. Set $A = 2^{1/\alpha}$: then $(Ax)^{-\alpha} = A^{-\alpha}x^{-\alpha} = 2^{-1}x^{-\alpha}$, not quite matching $2^{-\alpha}x^{-\alpha}$ unless $\alpha = 1$. The correct identification: $e^{-2^{-\alpha}x^{-\alpha}} = \Phi_\alpha(2x)$ since $\Phi_\alpha(2x) = e^{-(2x)^{-\alpha}} = e^{-2^{-\alpha}x^{-\alpha}}$. ✓ So $F^n(a'_n x) \to \Phi_\alpha(2x) = \Phi_\alpha(Ax + B)$ with $A = 2, B = 0$ — confirming the convergence-of-types theorem with $A = a_n/a'_n \to 1/2$ (consistent, since $a'_n = 2a_n$ gives $a_n/a'_n = 1/2$) and $B = 0$. **The "different" limit $\Phi_\alpha(2x)$ is simply Φ_α with**

a rescaled argument — the same Fréchet *type* (same α), differing only by the scale parameter, exactly as the convergence-of-types theorem guarantees. This is why, in practice, one is free to report $\hat{\sigma}_{\text{GEV}}$ (Section 1.3) in whatever units are convenient — the underlying shape parameter $\xi = 1/\alpha$, the quantity of real scientific interest, is invariant to such rescalings.

1.3 The Generalized Extreme Value (GEV) Distribution

Overview

The three families of Section 1.2 can be unified into a single parametric family, the Generalized Extreme Value (GEV) distribution, with a single shape parameter ξ that continuously interpolates between Weibull ($\xi < 0$), Gumbel ($\xi = 0$), and Fréchet ($\xi > 0$). This unification is not merely notational convenience: it is the basis of all modern statistical estimation in extreme value theory, since it allows ξ to be estimated from data without first deciding a priori which of the three "types" applies.

Definition (GEV distribution)

The Generalized Extreme Value distribution with location $\mu \in \mathbb{R}$, scale $\sigma > 0$, and shape $\xi \in \mathbb{R}$ has CDF

$$G_{\xi,\mu,\sigma}(x) = \begin{cases} \exp\left\{-\left[1+\xi\left(\dfrac{x-\mu}{\sigma}\right)\right]^{-1/\xi}\right\}, & \xi \neq 0, \quad 1+\xi\frac{x-\mu}{\sigma} > 0, \\ \exp\left\{-\exp\left(-\dfrac{x-\mu}{\sigma}\right)\right\}, & \xi = 0, \quad x \in \mathbb{R}. \end{cases}$$

The support is $x > \mu - \sigma/\xi$ if $\xi > 0$, $x < \mu - \sigma/\xi$ if $\xi < 0$, and all of $\mathbb{R}$ if $\xi = 0$.

Recovering the three types. Setting $\mu = 0, \sigma = 1$ and writing $\xi = 1/\alpha$ for $\xi > 0$:

$$G_{\xi>0}(x) = \exp\left\{-\left(1+\frac{x}{\alpha}\right)^{-\alpha}\right\} \xrightarrow{\text{sub } y=1+x/\alpha} \exp(-y^{-\alpha}) = \Phi_\alpha(y),$$

recovering the Fréchet form after the affine change of variable $y = 1 + \xi x$. Similarly $\xi < 0$ (writing $\xi = -1/\alpha$) recovers the Weibull form Ψ_α, and the limit $\xi \to 0$ recovers the Gumbel form Λ via the standard limit $(1+\xi x)^{-1/\xi} \to e^{-x}$.

The shape parameter ξ and tail heaviness. The parameter ξ (sometimes called the *extreme value index* or *tail index*, though the latter term is also used with a different normalization, $\alpha = 1/\xi$ for $\xi > 0$) governs the existence of moments of the limiting distribution:

ξ	Tail type	Moments
$\xi < 0$	Weibull-type, bounded support, upper endpoint $\mu - \sigma/\xi$	all moments finite
$\xi = 0$	Gumbel-type, exponential tail, unbounded support	all moments finite
$0 < \xi < 1$	Fréchet-type, polynomial tail $\sim x^{-1/\xi}$	mean finite, higher moments may d
$\xi \geq 1$	Fréchet-type, very heavy tail	mean itself infinite ($\xi \geq 1$)

More precisely, for $\xi > 0$, the k-th moment of $G_{\xi,\mu,\sigma}$ is finite if and only if $k < 1/\xi$. A value of $\xi = 0.5$, often found in financial loss data, implies that the limiting GEV has finite mean and finite variance but infinite third moment (no finite skewness) — a sobering fact for any risk model relying on higher moments.

Example — Moments of the GEV distribution

For $\xi \neq 0$ and $\xi < 1$, the mean of the standard ($\mu = 0, \sigma = 1$) GEV is

$$\mathbb{E}[X] = \frac{g_1 - 1}{\xi}, \qquad g_k \triangleq \Gamma(1 - k\xi),$$

provided $\xi < 1/k$ for g_k to be finite (recall $\Gamma(z)$ diverges at non-positive integers and $\Gamma(1-k\xi)$ requires $1 - k\xi > 0$). The variance is

$$\mathrm{Var}(X) = \frac{g_2 - g_1^2}{\xi^2}, \qquad \text{finite iff } \xi < \tfrac{1}{2}.$$

For $\xi = 0$ (Gumbel), $\mathbb{E}[X] = \gamma \approx 0.5772$ (the Euler–Mascheroni constant) and $\mathrm{Var}(X) = \pi^2/6$.

Numerical illustration. Take $\xi = 0.3$ (moderately heavy Fréchet-type tail, typical of insurance claim severities). Then:

$$g_1 = \Gamma(0.7) \approx 1.2981, \qquad g_2 = \Gamma(0.4) \approx 2.2182.$$

$$\mathbb{E}[X] = \frac{1.2981 - 1}{0.3} \approx 0.9937, \qquad \mathrm{Var}(X) = \frac{2.2182 - 1.2981^2}{0.3^2} = \frac{2.2182 - 1.6851}{0.09} \approx 5.924.$$

Note $\xi = 0.3 < 1/3$, so the third moment ($k = 3$ requires $\xi < 1/3 \approx 0.333$) is *just barely* finite; for $\xi = 0.35$ it would already diverge. This illustrates how sensitively the existence of higher moments depends on ξ, and why estimating ξ accurately (Section 1.5, Chapter 2) is of more than academic interest.

The GEV density. Differentiating the CDF gives, for $\xi \neq 0$ and $1 + \xi(x-\mu)/\sigma > 0$,

$$g_{\xi,\mu,\sigma}(x) = \frac{1}{\sigma}\left[1 + \xi\left(\frac{x-\mu}{\sigma}\right)\right]^{-1/\xi - 1} \exp\left\{-\left[1 + \xi\left(\frac{x-\mu}{\sigma}\right)\right]^{-1/\xi}\right\}.$$

Quantiles and return levels. Inverting $G_{\xi,\mu,\sigma}(x) = p$ gives the *quantile function*

$$x_p = \mu - \frac{\sigma}{\xi}\left[1 - (-\log p)^{-\xi}\right], \qquad \xi \neq 0,$$

$$x_p = \mu - \sigma \log(-\log p), \qquad \xi = 0.$$

These quantile expressions are the basis of *return level* estimation (Section 2.5): the level x_p exceeded by the annual maximum with probability $1 - p$ in any given year corresponds to a *return period* $T = 1/(1-p)$ years.

Example — Worked example: GEV fit and a 100-year return level

Suppose annual maximum river discharge (in m^3/s) has been fitted to a GEV with $\hat{\mu} = 320$, $\hat{\sigma} = 70$, $\hat{\xi} = 0.15$. The 100-year return level corresponds to $p = 1 - 1/100 = 0.99$:

$$-\log(0.99) \approx 0.01005, \qquad (-\log p)^{-\xi} = (0.01005)^{-0.15}.$$

Compute $(0.01005)^{-0.15} = \exp(-0.15 \ln 0.01005) = \exp(-0.15 \times (-4.600)) = \exp(0.690) \approx 1.994$.

$$x_{0.99} = 320 - \frac{70}{0.15}\left[1 - 1.994\right] = 320 - 466.67 \times (-0.994) \approx 320 + 463.9 \approx 783.9.$$

So the model predicts a discharge of roughly $784\,\mathrm{m}^3/\mathrm{s}$ is exceeded on average once every 100 years. Note the strong sensitivity to ξ: with $\xi = 0$ (Gumbel) instead,

$$x_{0.99} = \mu - \sigma \log(-\log 0.99) = 320 - 70 \times \log(0.01005) = 320 - 70 \times (-4.600) \approx 320 + 322 = 642,$$

a return level over 140 units lower. The choice between $\xi = 0$ and $\xi = 0.15$ — which may be statistically difficult to distinguish from a few decades of data — changes the 100-year design flood estimate by nearly 20%. This is a recurring theme: *shape-parameter uncertainty dominates extrapolation uncertainty* in extreme value applications.

The GEV as a likelihood-based model. In practice, ξ, μ, σ are estimated from data — typically a sequence of block maxima (Section 1.5) — by maximum likelihood, method of moments, or probability-weighted moments (Chapter 2). Maximum likelihood for the GEV is regular (asymptotically normal estimators with the usual $\sqrt{n}$ rate) only for $\xi > -1/2$; for $\xi < -1/2$ the standard regularity conditions fail because the support endpoint depends on the parameters themselves. Fortunately, $\xi < -1/2$ is rarely encountered in practice (it corresponds to extremely short-tailed, bounded phenomena), but the boundary case is a reminder that extreme value statistics, unlike most classical statistics, routinely operates near the edge of standard asymptotic theory.

The GEV as an exponential family — and why it isn't quite one. It is natural to ask whether the GEV, like the Normal or Gamma distributions, belongs to the exponential family of distributions — a classification with major implications for the existence of sufficient statistics and conjugate priors (Appendix K.1). The answer is subtle: **for fixed ξ, the GEV with varying (μ, σ) is *not* an exponential family** (the support depends on the parameters in a way incompatible with the exponential family's fixed-support requirement, except in the limiting Gumbel case $\xi = 0$, where the support is all of $\mathbb{R}$ regardless of (μ, σ)).

The Gumbel case as an exponential family

For $\xi = 0$, the GEV density is $g_{0,\mu,\sigma}(x) = \frac{1}{\sigma} \exp\left(-\frac{x-\mu}{\sigma} - e^{-(x-\mu)/\sigma}\right)$. Reparametrizing via $z = e^{-(x-\mu)/\sigma}$ (so $z \sim \mathrm{Exp}(1)$ when $X \sim \mathrm{Gumbel}(\mu, \sigma)$ — the Gumbel distribution is, by construction, the distribution of $\mu - \sigma \log Z$ for $Z \sim \mathrm{Exp}(1)$), one sees the Gumbel distribution is a location-scale transform of the (one-parameter exponential-family) log-Exponential distribution. **For $\xi \neq 0$, no such reduction to a fixed-support exponential family is possible** — the GEV for $\xi \neq 0$ is a genuinely non-exponential-family model, a fact with concrete consequences: there is no finite-dimensional sufficient statistic for (ξ, μ, σ) (other than the full order statistics themselves), unlike, say, the Normal distribution where $(\bar{X}, S^2)$ is sufficient for (μ, σ^2) regardless of n. This is part of why GEV/GPD inference (Sections 1.5, 2.4) requires numerical optimization of the full likelihood, with no shortcut via low-dimensional summary statistics — in marked contrast to, e.g., Gaussian-model inference where $(\bar{X}, S^2)$ suffices.

Skewness and kurtosis as functions of ξ: a fuller picture. Section 1.3 computed mean and variance via $g_1 = \Gamma(1-\xi)$, $g_2 = \Gamma(1-2\xi)$. The skewness requires $g_3 = \Gamma(1-3\xi)$ (finite for $\xi < 1/3$) and is

$$\mathrm{Skew}(X) = \mathrm{sgn}(\xi)\, \frac{g_3 - 3g_1 g_2 + 2g_1^3}{(g_2 - g_1^2)^{3/2}}.$$

Example — Numerical skewness across the ξ range

ξ	$g_1 = \Gamma(1-\xi)$	$g_2 = \Gamma(1-2\xi)$	Skewness
-0.2	$\Gamma(1.2) \approx 0.9182$	$\Gamma(1.4) \approx 0.8873$	≈ -0.41
0 (Gumbel)	1	1	1.1396 (fixed)
0.2	$\Gamma(0.8) \approx 1.1642$	$\Gamma(0.6) \approx 1.4892$	≈ 2.43
0.3	$\Gamma(0.7) \approx 1.2981$	$\Gamma(0.4) \approx 2.2182$	≈ 5.92

The Gumbel skewness (1.1396) is a *fixed constant*, independent of μ, σ — a useful diagnostic: an empirical sample of block maxima with skewness very different from 1.1396 suggests $\xi \neq 0$, and the *direction* (skewness > 1.1396 suggests $\xi > 0$; < 1.1396, including negative values, suggests $\xi < 0$) provides a quick, assumption-light first indication of the likely sign of ξ before any formal fitting — though, as Section 1.5 emphasized, sample skewness for small m is itself a high-variance statistic, and this diagnostic is best used as a rough prior check rather than a substitute for likelihood-based inference. Note also the extreme sensitivity near $\xi = 0.3$: skewness ≈ 5.92, already quite large, growing without bound as $\xi \to 1/3$ — a graphical illustration of how rapidly the GEV's shape changes in the Fréchet regime as ξ approaches successive moment-existence boundaries ($1/3$ for skewness, $1/2$ for variance as seen earlier, 1 for the mean).

1.4 Domains of Attraction and the Max-Stable Property

Overview

Having identified the three (unified: one-parameter GEV family of) possible limits, we now characterize precisely *which* distributions F converge to which limit — the theory of *domains of attraction*. The key tool is von Mises' sufficient conditions and the notion of *regularly varying* tails (developed fully in Chapter 3), together with the auxiliary function formalism that makes precise the heuristic "exponential vs. power-law vs. bounded" classification of Section 1.2.

The domain of attraction. We say F is in the *domain of attraction* of G_ξ (the GEV with shape ξ), written $F \in D(G_\xi)$, if there exist $a_n > 0, b_n \in \mathbb{R}$ with

$$F^n(a_n x + b_n) \to G_\xi(x) \quad \text{for all } x.$$

Domains of attraction partition (most of) the space of distributions into three classes corresponding to $\xi > 0$, $\xi = 0$, $\xi < 0$.

Fréchet domain ($\xi = 1/\alpha > 0$): regular variation. $F \in D(\Phi_\alpha)$ if and only if $\overline{F}(x) = 1 - F(x)$ is *regularly varying* with index $-\alpha$:

$$\overline{F}(x) = x^{-\alpha} L(x)$$

for some *slowly varying* function L, i.e. $\lim_{x\to\infty} L(tx)/L(x) = 1$ for all $t > 0$ (Chapter 3 develops this notion in full generality; here we use it as a black box). Equivalently, $\overline{F}(tx)/\overline{F}(x) \to t^{-\alpha}$ as $x \to \infty$, for every $t > 0$. The normalizing constants can be taken as $b_n = 0$ and $a_n = F^{-1}(1 - 1/n)$ (the $(1-1/n)$-quantile).

Weibull domain ($\xi = -1/\alpha < 0$). $F \in D(\Psi_\alpha)$ if and only if $x^* < \infty$ and the "reflected" survival function $\overline{F}(x^* - 1/x)$ is regularly varying with index $-\alpha$ as $x \to \infty$; equivalently $\overline{F}(x^* - x) = x^\alpha \tilde{L}(1/x)$ for slowly varying $\tilde{L}$. Intuitively, F approaches its finite upper endpoint x^* *polynomially*.

Gumbel domain ($\xi = 0$): von Mises' condition. This is the most heterogeneous domain, containing distributions with exponential, sub-exponential (e.g. log-normal), and even some moderately heavy tails (provided all moments are finite). A practically very useful *sufficient* condition, due to von Mises, is: if F is twice differentiable near x^* (possibly $+\infty$) with $f = F' > 0$, and

$$\lim_{x \to x^*} \frac{d}{dx}\left[\frac{\overline{F}(x)}{f(x)}\right] = 0,$$

then $F \in D(\Lambda)$ (Gumbel). The quantity $\overline{F}(x)/f(x)$ is called the *hazard reciprocal* or, in reliability terms (Volume II), the *mean residual life* of an exponential-type tail; the von Mises condition says this quantity varies slowly — changes negligibly relative to itself — near x^*.

Example — Verifying domain membership via von Mises' condition

Normal distribution. $\overline{F}(x) \sim \dfrac{\phi(x)}{x}$ as $x \to \infty$ (Mills' ratio), where ϕ is the standard normal density. Then

$$\frac{\overline{F}(x)}{f(x)} \sim \frac{\phi(x)/x}{\phi(x)} = \frac{1}{x} \to 0,$$

and

$$\frac{d}{dx}\left[\frac{1}{x}\right] = -\frac{1}{x^2} \to 0.$$

The von Mises condition holds: the Normal distribution is in the Gumbel domain, with norming constants $a_n = (2\log n)^{-1/2}$ and $b_n = (2\log n)^{1/2} - \frac{\log\log n + \log(4\pi)}{2(2\log n)^{1/2}}$ (a classical but somewhat unwieldy formula, reflecting the famously slow convergence of Gaussian maxima to their limit law — a cautionary fact for any naive application of asymptotic extreme value theory to Gaussian-modelled risks with moderate sample sizes).

Exponential distribution. $\overline{F}(x) = e^{-\lambda x}$, $f(x) = \lambda e^{-\lambda x}$, so $\overline{F}(x)/f(x) = 1/\lambda$, a constant, whose derivative is identically 0. Von Mises' condition holds trivially — in fact *exactly*, not just asymptotically, consistent with the exact finite-n result of Section 1.2(a).

Log-normal distribution. Although log-normal tails decay slower than any exponential (the density behaves like $\exp(-(\log x)^2/2\sigma^2)/x$, which is "sub-exponential"), one can show $\overline{F}(x)/f(x) \sim \sigma^2 x/\log x \to \infty$, but its derivative $\to 0$ — the von Mises condition is satisfied in the limit, placing log-normal in the Gumbel domain despite its notoriously heavy-looking tail. This is a famous source of confusion in risk modelling: *log-normal losses, despite "looking" heavy-tailed in finite samples, are technically light-tailed (Gumbel domain) in the extreme value sense*, with consequences for tail risk estimates that we revisit in Chapter 3.

Domains of attraction and stability under maxima. The max-stability property $G(x)^n = G(\alpha_n x + \beta_n)$ has a striking practical corollary: if losses $X_1, \ldots, X_n$ are i.i.d. Fréchet(α) *exactly* (not just asymptotically), then $\max(X_1, \ldots, X_n)$ is *again* Fréchet(α) after rescaling by $n^{1/\alpha}$ — the family is closed under taking maxima. This self-similarity is the deep reason power laws recur across scales in natural and economic systems: a process whose extremes are governed by a Fréchet (equivalently, Pareto-tailed) mechanism looks "the same" (up to rescaling) whether one examines daily, monthly, or yearly maxima.

> **Practical implication.** The domain of attraction of a loss distribution determines not just the *shape* of extreme-value models but the entire *philosophy* of risk extrapolation.

In the Fréchet domain, the heaviness of the tail (parameter α) is preserved under aggregation and under taking maxima — "one big jump" dominates sums and maxima alike (formalized in Chapter 3's principle of a single big jump). In the Gumbel domain, by contrast, sums are dominated by the CLT (many moderate contributions) while maxima still follow GEV theory — the two types of extremal behaviour (sum vs. maximum) can diverge sharply, a distinction of central importance when choosing between Value-at-Risk-type measures (based on quantiles, related to maxima/order statistics) and measures based on aggregate loss distributions (Chapter 4).

Second-order regular variation and rates of convergence. The domain of attraction characterizes *whether* $F^n(a_n x + b_n) \to G_\xi(x)$, but says nothing about the *rate*. As the Normal example shows, convergence can be excruciatingly slow (logarithmic in n). The theory of *second-order regular variation* introduces an auxiliary parameter $\rho \leq 0$ controlling the rate of convergence to the first-order limit, of the form

$$\frac{F^n(a_n x + b_n) - G_\xi(x)}{A(n)} \to H_{\xi,\rho}(x)$$

for some second-order limit function H and a rate sequence $A(n) \to 0$. This refinement underlies bias-corrected estimators of ξ (Section 2.4) and is essential whenever extreme value methods are applied to moderate (rather than astronomically large) sample sizes — which is to say, in essentially all real applications.

A unified view via the auxiliary function. The three domain-of-attraction characterizations (Fréchet via regular variation of $\overline{F}$; Weibull via regular variation of the reflected tail; Gumbel via the von Mises condition) can be unified through the *auxiliary function* $a(t)$, defined via

$$a(t) = \frac{\int_t^{x^*} \overline{F}(s)\, ds}{\overline{F}(t)}$$

(for $x^* = +\infty$; an analogous definition handles finite x^*). This is precisely $\mathbb{E}[X - t \mid X > t]$, the **mean excess function** already encountered for the GPD in Section 2.1 — but here defined for an *arbitrary* F, not just the GPD, and serving a unifying theoretical role rather than (yet) a statistical one.

Theorem (Unified domain-of-attraction criterion via $a(t)$)

$F \in D(G_\xi)$ for some $\xi \in \mathbb{R}$ if and only if there exists a function $a(t) > 0$ (which may be taken as the auxiliary function above) such that

$$\lim_{t \to x^*} \frac{\overline{F}(t + x a(t))}{\overline{F}(t)} = (1 + \xi x)_+^{-1/\xi}$$

for all x with $1 + \xi x > 0$ (with the right side interpreted as e^{-x} for $\xi = 0$, by the usual limiting convention).

This single criterion — with $\xi > 0$, $\xi = 0$, $\xi < 0$ as special cases — subsumes the Fréchet (regular variation), Gumbel (von Mises), and Weibull characterizations of Sections 1.2/1.4 *simultaneously*: regular variation of $\overline{F}$ with index $-\alpha$ corresponds to $a(t) = t/\alpha$ (so $a(t)/t \to 1/\alpha$, a constant ratio — the auxiliary function grows linearly with t); the von Mises condition corresponds to $a(t) \to$ some function with $a'(t) \to 0$ (the auxiliary function varies slowly relative to t, in a precise sense); and the Weibull case corresponds to $a(t) \to 0$ as $t \to x^* < \infty$ in a specific manner.

Example — The auxiliary function as a direct link to GPD scale

Recall from Section 2.1/2.2 that, for $F \in D(G_\xi)$, the GPD scale parameter $\beta(u)$ in the PBdH theorem satisfies $\beta(u) \approx a(u) \times \xi$ for $\xi \neq 0$ (more precisely, $\beta(u)/a(u) \to \xi$ as $u \to x^*$ for $\xi \neq 0$, while for $\xi = 0$, $\beta(u)/a(u) \to 1$). **The auxiliary function $a(t)$ — defined purely in terms of the original F, with no reference to any limiting distribution — is thus, asymptotically, (proportional to) the GPD scale parameter at threshold t.** This provides a direct, non-asymptotic-distribution-based way to estimate $\beta(u)$ for finite u: simply estimate the mean excess function $\hat{e}(u) = a(u)$ empirically (exactly as in Section 2.3's mean excess plot), and the GPD scale at that threshold is (asymptotically, for $\xi \neq 0$) $\hat{\beta}(u) \approx \xi\, \hat{e}(u)$ — connecting the threshold-selection diagnostic of Section 2.3 directly back to the abstract domain-of-attraction theory of this section, and underscoring that the mean excess function is not merely a useful empirical heuristic but is, in a precise sense, *the* object that characterizes domains of attraction at the most fundamental level.

1.5 Block Maxima Method: Estimation and Diagnostics

Overview

We turn from probability theory to statistics: given a finite time series of observations, how does one estimate the GEV parameters (ξ, μ, σ) governing its extremes? The *block maxima* method — partition the data into blocks, extract the maximum of each, and fit a GEV to the resulting sample of maxima — is the oldest and most intuitive approach, directly motivated by the Fisher–Tippett–Gnedenko theorem. We cover maximum likelihood estimation, diagnostic plots, and the crucial practical question of block-size selection.

The method. Given observations $X_1, \dots, X_N$ (e.g. daily values over many years), partition the index set $\{1, \dots, N\}$ into m contiguous blocks of size $n = N/m$ (e.g. $n = 365$ for annual blocks of daily data), and define

$$M_j = \max\{X_i : i \text{ in block } j\}, \qquad j = 1, \dots, m.$$

By the FTG theorem, for n large, $M_j \overset{approx}{\sim} G_{\xi,\mu,\sigma}$ (GEV), and one fits this distribution to the sample $M_1, \dots, M_m$ by maximum likelihood (ML), treating $M_1, \dots, M_m$ as approximately i.i.d.

Maximum likelihood estimation. The log-likelihood for (ξ, μ, σ) given m block maxima $M_1, \dots, M_m$, for $\xi \neq 0$, is

$$\ell(\xi, \mu, \sigma) = -m \log \sigma - \left(1 + \frac{1}{\xi}\right) \sum_{j=1}^{m} \log z_j - \sum_{j=1}^{m} z_j^{-1/\xi},$$

where $z_j = 1 + \xi \dfrac{M_j - \mu}{\sigma}$, valid provided $z_j > 0$ for all j (else the likelihood is 0, i.e. $\log \ell = -\infty$). This constraint — that the support of the fitted GEV must contain all observed block maxima — means the likelihood surface is genuinely constrained, and standard unconstrained optimizers must be used with care (or the constraint built in via reparametrization, e.g. optimizing over $z_j > 0$ directly).

For $\xi = 0$ (Gumbel), the log-likelihood simplifies to

$$\ell(0, \mu, \sigma) = -m \log \sigma - \sum_{j=1}^{m} \frac{M_j - \mu}{\sigma} - \sum_{j=1}^{m} \exp\left(-\frac{M_j - \mu}{\sigma}\right).$$

Asymptotic properties of the MLE. For $\xi > -1/2$, the GEV satisfies the regularity conditions for maximum likelihood, and $(\hat{\xi}, \hat{\mu}, \hat{\sigma})$ is asymptotically normal:

$$\sqrt{m}\left(\begin{pmatrix}\hat{\xi}\\ \hat{\mu}\\ \hat{\sigma}\end{pmatrix} - \begin{pmatrix}\xi\\ \mu\\ \sigma\end{pmatrix}\right) \xrightarrow{d} \mathcal{N}(\mathbf{0}, I(\xi, \mu, \sigma)^{-1}),$$

where I is the Fisher information matrix. Closed-form expressions for I^{-1} exist (Prescott & Walden, 1980) but are complicated; in practice, the inverse observed information matrix at the MLE is used to construct confidence intervals.

Example — Worked numerical example: fitting a GEV by ML

Suppose $m = 50$ years of annual maximum daily rainfall (mm) are summarized by sample mean $\bar{M} = 85.2$, sample standard deviation $s_M = 18.4$, and sample skewness $\hat{\gamma}_1 = 0.62$.
A quick method-of-moments starting point uses the relations (for ξ near 0, expanding to first order):

$$\mathbb{E}[M] \approx \mu + \sigma\gamma, \qquad \mathrm{Var}(M) \approx \sigma^2 \frac{\pi^2}{6}, \qquad \text{skewness} \approx 1.1396\,\xi \quad (\text{for small } \xi > 0),$$

where $\gamma \approx 0.5772$. From the skewness, $\hat{\xi}^{(0)} \approx 0.62/1.1396 \approx 0.544$ — but this exceeds $1/2$, signalling that the variance of the GEV itself may be infinite at this shape, a red flag suggesting either a genuinely very heavy-tailed process or a small-sample skewness estimate contaminated by one or two extreme block maxima (a common occurrence: $\hat{\gamma}_1$ has very high variance for $m = 50$).
A full numerical ML fit (which we summarize without re-deriving the optimization) yields, say,

$$\hat{\xi} = 0.21 \ (\text{SE } 0.13), \qquad \hat{\mu} = 76.4 \ (\text{SE } 2.6), \qquad \hat{\sigma} = 15.8 \ (\text{SE } 2.1).$$

The standard error on $\hat{\xi}$ (0.13) is large relative to its value (0.21): a 95% confidence interval $\hat{\xi} \pm 1.96 \times 0.13 \approx (-0.045, 0.465)$ comfortably straddles 0 (Gumbel) and approaches 0.5 (the threshold for finite variance). This is entirely typical: *with $m = 50$ block maxima, the shape parameter is very poorly constrained*, and conclusions about whether the underlying process is "heavy-tailed" (Fréchet, $\xi > 0$) versus "light-tailed" (Gumbel, $\xi = 0$) are often not statistically resolvable from a single century or so of annual data — a humbling but important practical lesson.

Diagnostic plots. Three standard diagnostics assess GEV fit:

(1) *Probability plot.* Plot empirical probabilities $j/(m+1)$ (or other plotting positions) against $\hat{G}(M_{(j)})$ for the ordered block maxima $M_{(1)} \le \ldots \le M_{(m)}$; points should lie near the diagonal $y = x$.

(2) *Quantile (Q-Q) plot.* Plot $M_{(j)}$ against $\hat{G}^{-1}(j/(m+1))$; again, points should lie near $y = x$. Q-Q plots are generally more informative than probability plots for assessing tail fit, since

deviations in the extreme tail are compressed near 0 or 1 in probability scale but spread out in quantile scale.

(3) *Return level plot.* Plot the fitted return level x_p (Section 1.3) against $-\log(-\log p)$ (so that a Gumbel fit, $\xi = 0$, appears as a straight line), with empirical points and confidence bands overlaid; curvature reveals departures from $\xi = 0$.

The block-size trade-off. Block maxima estimation faces a classic bias-variance trade-off in the choice of block size n (equivalently, the choice of $m = N/n$ given fixed total data N):

- *Larger* n (fewer, larger blocks $\Rightarrow$ smaller m): the GEV approximation $F^n \approx G_\xi$ becomes more accurate (lower *bias*), but fewer block maxima are available for fitting, increasing *variance* of $(\hat{\xi}, \hat{\mu}, \hat{\sigma})$.
- *Smaller* n (more, smaller blocks $\Rightarrow$ larger m): more data points for fitting (lower variance), but the GEV approximation is less accurate for small n (higher bias), since F^n has not yet "converged" to its limiting shape.

In practice, the choice of n is often dictated by natural periodicities (e.g. annual blocks for hydrological/meteorological data, to respect seasonality), but the underlying statistical trade-off is real and is the principal motivation for the *threshold-based* (peaks-over-threshold) methods of Chapter 2, which make more efficient use of the data by using *all* sufficiently extreme observations rather than only one per block.

Profile likelihood for ξ. Because $\hat{\xi}$ is typically the parameter of greatest practical interest (it governs tail heaviness and the existence of moments, Section 1.3) and is often poorly constrained (as in the worked example above), it is standard practice to report a *profile likelihood* confidence interval for ξ rather than relying on the (potentially poor, for small m) normal approximation. The profile log-likelihood is

$$\ell_p(\xi) = \max_{\mu,\sigma} \ell(\xi, \mu, \sigma),$$

and an (asymptotically) $100(1-\alpha)\%$ confidence interval for ξ is $\{\xi : 2[\ell(\hat{\xi}, \hat{\mu}, \hat{\sigma}) - \ell_p(\xi)] \leq \chi^2_{1,1-\alpha}\}$, by Wilks' theorem. Profile likelihood intervals are typically asymmetric — often extending further into the Fréchet ($\xi > 0$) region than the Weibull ($\xi < 0$) region for loss data — reflecting the genuine asymmetry of statistical information about tail heaviness versus tail boundedness.

PART 2

Threshold Methods and the Generalized Pareto Distribution

Peaks-over-threshold methodology, the Pickands–Balkema–de Haan theorem, and practical estimation of tail behaviour.

The information matrix and why ξ is hard to estimate: a closer look. Section 1.5 reported that $\mathrm{SE}(\hat{\xi}) \approx 0.13$ for $m = 50$, larger (relatively) than the standard errors for μ, σ. This is not an artifact of the particular numbers chosen but reflects a structural feature of the GEV's Fisher information matrix, which we now examine.

The GEV Fisher information (standardized form)

For the standard GEV ($\mu = 0, \sigma = 1$), the Fisher information matrix $I(\xi)$ for (ξ, μ, σ) (evaluated at $\mu = 0, \sigma = 1$, by location-scale equivariance the general case follows by

simple rescaling) has, as $\xi \to 0$, the entries

$$I_{\xi\xi}(\xi) \to \frac{\pi^2}{6} + (1-\gamma)^2 + 1 \approx 2.4222$$

(where $\gamma \approx 0.5772$ is the Euler–Mascheroni constant), a *finite, modest-sized* constant. By contrast, the entries $I_{\mu\mu}$ and $I_{\sigma\sigma}$ are also $O(1)$ constants of comparable or larger magnitude as $\xi \to 0$. The seemingly larger *relative* uncertainty in $\hat{\xi}$ does not come from $I_{\xi\xi}$ being small in absolute terms, but from the *off-diagonal* correlations: $I^{-1}_{\xi\xi}$ (the asymptotic variance of $\hat{\xi}$, i.e. the (ξ,ξ) entry of the *inverse* information matrix) is inflated by the substantial correlations between $\hat{\xi}$ and $\hat{\sigma}$ (typically $\mathrm{Corr}(\hat{\xi},\hat{\sigma})$ in the range -0.4 to -0.6 for ξ near 0), a correlation structure absent (or much weaker) for $(\hat{\mu},\hat{\sigma})$.

Why the ξ-σ correlation arises. Intuitively, ξ and σ both affect the *spread* of the distribution — a larger σ stretches the distribution uniformly, while a larger ξ stretches it more in the tail than in the bulk. For a finite, moderately-sized sample, these two effects are difficult to disentangle: an apparently "spread out" sample of block maxima could be explained by large σ with small ξ, or by smaller σ with larger ξ (concentrating the "extra spread" in the tail) — the likelihood surface has a "ridge" along combinations of (ξ,σ) that produce similar overall spread, and this ridge is what produces the negative correlation and the resulting inflation of $\mathrm{Var}(\hat{\xi})$ via the matrix inversion.

Example — Numerical illustration of the information-matrix inflation

Suppose, hypothetically, $(\hat{\xi},\hat{\sigma})$ were *uncorrelated*, each with the same marginal variance as in the actual (correlated) case. If $I^{-1}_{\xi\xi} \approx 0.094^2 \times m = 0.0088 \times 50 \approx 0.441$ (the per-observation asymptotic variance, from Section 1.5's $\mathrm{SE}(\hat{\xi}) = 0.094$ for $k = 180$... wait, that was the GPD example; for the GEV $m = 50$ example, $\mathrm{SE} = 0.13$, so per-observation variance $\approx 0.13^2 \times 50 \approx 0.845$). If the off-diagonal correlation $\rho_{\xi\sigma} \approx -0.5$ were instead 0, the resulting $I^{-1}_{\xi\xi}$ (which, for a 2×2 block, is $I_{\xi\xi}/(I_{\xi\xi}I_{\sigma\sigma} - I^2_{\xi\sigma})$ if we ignore μ for simplicity, i.e. inversely related to $1-\rho^2_{\xi\sigma}$ times the marginal) would be smaller by a factor of roughly $1-\rho^2_{\xi\sigma} = 1 - 0.25 = 0.75$ — i.e. $\mathrm{SE}(\hat{\xi})$ would be smaller by a factor of $\sqrt{0.75} \approx 0.866$, or about 13%. **Roughly 13% of the standard error on $\hat{\xi}$ in this illustration is attributable to the $\hat{\xi}$-$\hat{\sigma}$ correlation alone** — a correlation that is a structural feature of the GEV's geometry near $\xi = 0$, not a finite-sample artifact that vanishes with more data (the correlation $\rho_{\xi\sigma}$ converges to a fixed, non-zero limit as $m \to \infty$; it is the *variances* that shrink with m, not the correlation). This is part of why, as Appendix K.1 discussed, the choice of prior for ξ in a Bayesian analysis can interact non-trivially with the prior for σ (or β, for the GPD) — the parameters are not asymptotically independent, and a prior that treats them as independent may not faithfully represent genuine prior beliefs about the *joint* shape of the tail.

2.1 The Generalized Pareto Distribution (GPD)

Overview

The Generalized Pareto Distribution (GPD) is to *excesses over a threshold* what the GEV is to *block maxima*: the universal limiting law, this time for the conditional distribution of $X - u$ given $X > u$ as the threshold u approaches the right endpoint x^*. The GPD shares the same shape parameter ξ as the GEV — a deep and useful connection — and underlies the peaks-over-threshold (POT) approach developed throughout this chapter.

Definition (Generalized Pareto Distribution)

The GPD with scale $\beta > 0$ and shape $\xi \in \mathbb{R}$ has CDF

$$H_{\xi,\beta}(y) = \begin{cases} 1 - \left(1 + \xi \dfrac{y}{\beta}\right)^{-1/\xi}, & \xi \neq 0, \\ 1 - \exp(-y/\beta), & \xi = 0, \end{cases} \qquad y \geq 0,$$

with support $y \in [0, \infty)$ if $\xi \geq 0$, and $y \in [0, -\beta/\xi]$ if $\xi < 0$.

Special cases. For $\xi = 0$, $H_{0,\beta}$ is the ordinary Exponential(β) distribution. For $\xi > 0$, writing $\alpha = 1/\xi$ and $y = \beta(x/\beta - 1)$ shows $H_{\xi,\beta}$ is, up to a shift and scale, the classical Pareto distribution $\overline{H}(x) = (x/\beta)^{-\alpha}$ for $x \geq \beta$ — hence the name. For $\xi < 0$, $H_{\xi,\beta}$ has bounded support and is, up to reparametrization, a (negative-power, i.e. "Pareto type II / Lomax with negative shape") Beta-like distribution on a finite interval.

Survival function and hazard rate. The survival function $\overline{H}_{\xi,\beta}(y) = 1 - H_{\xi,\beta}(y) = (1 + \xi y/\beta)^{-1/\xi}$ (for $\xi \neq 0$) has hazard rate

$$h(y) = \frac{H'_{\xi,\beta}(y)}{\overline{H}_{\xi,\beta}(y)} = \frac{1}{\beta + \xi y}.$$

This is *linear* in y (for $\xi \neq 0$): increasing if $\xi < 0$ (bounded support, hazard $\to \infty$ as $y \to -\beta/\xi$), constant if $\xi = 0$ (the memoryless Exponential), and decreasing if $\xi > 0$ (heavy tail, hazard $\to 0$). This linearity of the hazard — or equivalently, of the *mean excess function* introduced below — is the GPD's defining and most practically useful characteristic, since it gives a simple graphical diagnostic.

Definition (Mean excess function)

For a random variable X with finite mean above u, the *mean excess function* (also: mean residual life) is

$$e(u) = \mathbb{E}[X - u \mid X > u].$$

Example — Mean excess function of the GPD

For $Y \sim H_{\xi,\beta}$ and any threshold $v \geq 0$ (with $v < -\beta/\xi$ if $\xi < 0$), using the fact that exceedances of a GPD over a higher threshold are again GPD-distributed with an updated scale (the key stability property derived next):

$$e(v) = \mathbb{E}[Y - v \mid Y > v] = \frac{\beta + \xi v}{1 - \xi}, \qquad \xi < 1.$$

This is *linear in* v with slope $\xi/(1-\xi)$ and intercept $\beta/(1-\xi)$. Three regimes:

- $\xi > 0$: $e(v)$ increases linearly in v — the mean excess grows as the threshold rises, the signature of a heavy (Fréchet-type) tail.
- $\xi = 0$: $e(v) = \beta$, constant — the memoryless Exponential case.
- $\xi < 0$: $e(v)$ decreases linearly, reaching $e(-\beta/\xi) = 0$ at the upper endpoint — the signature of bounded support.

This linear mean-excess plot is the basis of the diagnostic threshold-selection procedure of Section 2.3.

The threshold stability property. The GPD possesses a remarkable self-similarity under further thresholding: if $Y \sim H_{\xi,\beta}$, then for any $v \geq 0$ (within the support), the conditional excess $Y - v \mid Y > v$ is *again* GPD-distributed, with the *same* shape ξ but updated scale $\beta + \xi v$:

$$(Y - v \mid Y > v) \sim H_{\xi,\, \beta+\xi v}.$$

Proof. For $\xi \neq 0$,

$$P(Y - v > y \mid Y > v) = \frac{P(Y > v + y)}{P(Y > v)} = \frac{(1 + \xi(v + y)/\beta)^{-1/\xi}}{(1 + \xi v/\beta)^{-1/\xi}} = \left(1 + \frac{\xi y}{\beta + \xi v}\right)^{-1/\xi},$$

where the last step factors $1 + \xi(v+y)/\beta = (1 + \xi v/\beta)\left(1 + \frac{\xi y}{\beta+\xi v}\right)$, a direct algebraic identity. This is exactly $\overline{H}_{\xi,\beta+\xi v}(y)$. ■

This property is the GPD analogue of max-stability for the GEV, and it is the precise mathematical reason that, *if* excesses over some threshold u_0 follow a GPD(ξ, β_0), then excesses over any *higher* threshold $u > u_0$ also follow a GPD with the *same* ξ — a testable consequence used directly in threshold diagnostics (Section 2.3).

Connection to the exponential distribution and order statistics. The GPD with $\xi = 0$ is the Exponential distribution, which is the unique *memoryless* continuous distribution: $P(X > s + t \mid X > s) = P(X > t)$. The GPD generalizes this memorylessness to a one-parameter family of "threshold-stable" laws. This connects directly to the theory of order statistics of Section 1.1: just as the Exponential arises as the limit of rescaled minima of Uniform variables, the GPD arises as the limit of rescaled *excesses* for essentially any underlying distribution, by the Pickands–Balkema–de Haan theorem of the next section — a result of comparable universality to the FTG theorem itself.

The GPD as a limiting case of the Generalized Beta of the Second Kind. A useful perspective on the GPD's place within the broader landscape of heavy-tailed distributions comes from the *Generalized Beta Distribution of the Second Kind* (GB2), a four-parameter family with density

$$f_{\text{GB2}}(x) = \frac{|a| x^{ap-1}}{b^{ap} B(p,q)[1 + (x/b)^a]^{p+q}}, \qquad x > 0,$$

where $B(p, q)$ is the Beta function. The GB2 nests an enormous number of standard distributions as special or limiting cases: the Generalized Gamma ($q \to \infty$), the Burr Type XII / Singh–Maddala distribution ($p = 1$), the Lomax/Pareto Type II ($a = p = 1$), and — of direct relevance here — **the GPD arises as the limit $q \to \infty$ of the Burr distribution with appropriate rescaling**, situating the GPD as one vertex of this broader family.

Example — Why this matters: alternatives when the GPD fits poorly

If a mean-excess plot (Section 2.3) shows curvature inconsistent with the GPD's linear mean-excess property — e.g. a mean excess function that is concave (increasing but at a decreasing rate) rather than linear — this may indicate the true tail, while still heavy, has a more complex shape than the GPD captures. The Burr distribution (a GB2 special case with one fewer parameter than full GB2, hence more tractable) provides a natural one-parameter-richer alternative: its survival function $\overline{F}(x) = [1 + (x/b)^a]^{-q}$ behaves like $\overline{F}(x) \sim (x/b)^{-aq}$ for large x (so it is regularly varying with index $-aq$, same Fréchet-domain classification as a GPD with $\xi = 1/(aq)$) but with a different *rate of approach* to this asymptotic power-law behaviour, governed by a — giving the Burr distribution more flexibility to fit the "shoulder" of the distribution (the transition region between the bulk and the asymptotic tail) while preserving the same ultimate tail index. In practice, fitting a Burr distribution to the *entire* dataset (not just exceedances above a threshold) and comparing its implied tail index $\xi_{\text{Burr}} = 1/(aq)$ to the GPD-based $\hat{\xi}$ from Section 2.4 provides a useful cross-check: substantial disagreement suggests the threshold-based GPD approach may be sensitive to the specific threshold chosen (Section 2.3), since the Burr fit implicitly "smooths" across the threshold-selection decision by using all the data.

A note on the historical name "Generalized Pareto." The ordinary (Type I) Pareto distribution, $\overline{F}(x) = (x/x_{\min})^{-\alpha}$ for $x \geq x_{\min}$, corresponds to the GPD with $\xi = 1/\alpha$ *and* a specific relationship between β and the threshold $u = x_{\min}$ (namely $\beta = \xi u$, so that the GPD's lower endpoint coincides with $x_{\min}$ in original units — this is precisely the threshold-stability relation of Section 2.1's main text, with $u_0 = 0, \beta_0 = 0$ as a degenerate starting point). The GPD "generalizes" Pareto in two distinct directions simultaneously: it allows $\xi \leq 0$ (Exponential and bounded-support cases, not present in the classical Pareto family at all), and — even for $\xi > 0$ — it decouples the scale β from the threshold u, allowing the GPD to describe excesses above *any* threshold u (not just $u = 0$ as for the "standard" Pareto), which is precisely the threshold-stability property that makes the GPD useful for the POT framework of this chapter.

2.2 The Pickands–Balkema–de Haan Theorem

Overview

The Pickands–Balkema–de Haan (PBdH) theorem (1974/1975) is the second pillar of extreme value theory, complementing the Fisher–Tippett–Gnedenko theorem. While FTG concerns the limiting distribution of the *maximum*, PBdH concerns the limiting distribution of *excesses over a threshold*, and asserts that this limit is always a GPD — with shape parameter ξ exactly matching that of the GEV limit for the same underlying F. This shared ξ is the linchpin connecting block-maxima and threshold-based inference.

Theorem (Pickands–Balkema–de Haan)

$F \in D(G_\xi)$ (i.e. F is in the domain of attraction of the GEV with shape ξ, Section 1.4) *if and only if* there exists a positive function $\beta(u)$ such that

$$\lim_{u \to x^*} \sup_{0 \leq y < x^* - u} \left| F_u(y) - H_{\xi,\beta(u)}(y) \right| = 0,$$

where $F_u(y) = P(X - u \le y \mid X > u)$ is the conditional excess distribution above threshold u, and $H_{\xi,\beta}$ is the GPD of Section 2.1.

In words: *for any F in a domain of attraction, the distribution of excesses over a sufficiently high threshold u is well-approximated by a GPD with shape ξ equal to the GEV shape for F, and some scale $\beta(u)$ depending on u.*

Proof idea via the GEV connection. Suppose $F \in D(G_\xi)$, so $F^n(a_n x + b_n) \to G_\xi(x)$. Consider $u_n = a_n x + b_n$ for large x (i.e. $u_n \to x^*$), and write

$$n(1 - F(u_n)) \to -\log G_\xi(x) = (1 + \xi x)^{-1/\xi}.$$

Now consider an excess y above u_n, i.e. the point $u_n + a_n y' \approx a_n(x + y') + b_n$ for small y' representing an additional excess on the normalized scale. The ratio

$$\frac{1 - F(u_n + a_n y')}{1 - F(u_n)} \to \frac{(1 + \xi(x + y'))^{-1/\xi}}{(1 + \xi x)^{-1/\xi}} = \left(1 + \frac{\xi y'}{(1 + \xi x)/1}\right)^{-1/\xi} = \left(1 + \frac{\xi y'}{1 + \xi x}\right)^{-1/\xi},$$

using the same algebraic factoring identity as in Section 2.1's proof of threshold stability. The right-hand side is exactly $\overline{H}_{\xi,(1+\xi x)}(y')$ — a GPD survival function in y' with scale $\beta = 1 + \xi x$ (in the normalized units a_n). Translating back to the original scale gives $\beta(u) = a_n(1 + \xi x)$, completing the heuristic. A fully rigorous proof requires controlling the uniformity of these limits (the supremum in the theorem statement), which is where the technical content of regular variation theory (Chapter 3) enters. ■

The relationship between $\beta(u)$ and the GPD threshold-stability property. Combining PBdH with the threshold-stability property of Section 2.1, one finds that $\beta(u)$ is necessarily of the form

$$\beta(u) = \beta_0 + \xi(u - u_0)$$

for any reference threshold u_0 and corresponding $\beta_0 = \beta(u_0)$ — *linear in u*, with slope exactly ξ. This is not an additional assumption but a *consequence* of the GPD's threshold-stability: if excesses above u_0 are GPD(ξ, β_0), the theorem (applied iteratively, or via the exact identity of Section 2.1) shows excesses above any $u > u_0$ are GPD$(\xi, \beta_0 + \xi(u - u_0))$, exactly.

Example — Worked example: deriving GPD parameters from a known F

Let X have CDF $F(x) = 1 - x^{-2}e^{-1/x}$ for $x \ge 1$ (a distribution chosen to be analytically tractable but representative of mixed power-exponential tail behaviour). As $x \to \infty$, $e^{-1/x} \to 1$, so $\overline{F}(x) \sim x^{-2}$ — regularly varying with index -2, placing F in the Fréchet domain with $\xi = 1/2$ (Section 1.4).
For a high threshold u, the conditional excess distribution is

$$\overline{F}_u(y) = \frac{\overline{F}(u + y)}{\overline{F}(u)} = \frac{(u + y)^{-2}e^{-1/(u+y)}}{u^{-2}e^{-1/u}} \approx \left(\frac{u + y}{u}\right)^{-2} = \left(1 + \frac{y}{u}\right)^{-2}$$

as $u \to \infty$ (the exponential factors $\to 1$ and cancel to leading order). Comparing to $\overline{H}_{\xi,\beta}(y) = (1 + \xi y/\beta)^{-1/\xi}$ with $\xi = 1/2$:

$$\left(1 + \frac{y}{u}\right)^{-2} = \left(1 + \frac{(1/2)y}{u/2}\right)^{-1/(1/2)} \quad \Rightarrow \quad \beta(u) = u/2.$$

Indeed $\beta(u) = u/2 = \xi u$ is linear in u with slope $\xi = 1/2$ and intercept $\beta(0) = 0$, exactly as PBdH predicts. This explicit example illustrates how the abstract theorem cashes out into concrete, checkable formulas.

Practical consequence: the POT model. The PBdH theorem licenses the following statistical model, the basis of *peaks-over-threshold* (POT) analysis: choose a high threshold u; model the excesses $\{X_i - u : X_i > u\}$ as i.i.d. draws from $H_{\xi,\beta}$ for some (ξ, β); and separately model the *number* of exceedances (Section 2.3) — typically via a Poisson process for the exceedance times. The combination of a GPD for excess magnitudes and a Poisson process for exceedance times is called the *Poisson-GPD model*, and it is mathematically equivalent (in an appropriate limit) to the GEV block-maxima model, but typically makes more efficient use of data because it does not discard all but one observation per block.

> **The GEV–GPD shape correspondence is exact, not approximate, in the limit.** A common error is to fit a GEV to block maxima and a GPD to threshold excesses *independently* and be surprised when the two $\hat{\xi}$ estimates differ. PBdH guarantees these should converge to the *same* ξ as the block size / threshold $\to \infty$; finite-sample discrepancies reflect estimation variance and/or the different rates of convergence to the respective limits (block maxima typically converging more slowly than threshold excesses for a fixed amount of data, since each block discards all but its maximum). Reconciling GEV- and GPD-based estimates of ξ is a standard — and informative — diagnostic in applied extreme value analysis.

A converse perspective: when does PBdH *fail*? The PBdH theorem (Section 2.2) asserts that *if* $F \in D(G_\xi)$, then threshold excesses converge to a GPD. It is instructive to ask: are there distributions F for which *no* domain of attraction applies at all — i.e. $F \notin D(G_\xi)$ for any ξ — and what does PBdH say (or fail to say) in such cases?

Distributions outside any domain of attraction

A classical example is $\overline{F}(x) = \exp(-x(2 + \sin(\log x)))$ for $x \geq 1$ (a survival function that oscillates, via the $\sin(\log x)$ term, between faster and slower exponential-type decay as $x \to \infty$). One can show this F is *not* in any domain of attraction $D(G_\xi)$: the ratio $\overline{F}(tx)/\overline{F}(x)$ does not converge to a limit as $x \to \infty$ for fixed $t \neq 1$ (it oscillates), violating the regular-variation-type conditions (Section 1.4) needed for any of the three domains.

Consequence for POT in such cases. For such F, the conditional excess distribution $F_u(y) = P(X - u \leq y \mid X > u)$ does *not* converge to any GPD as $u \to \infty$ — it continues to oscillate between different "apparent" GPD shapes depending on u. In practice, this manifests as a Hill plot (Section 2.3) or mean-excess plot that, no matter how high the threshold, never stabilizes — it continues to drift or oscillate as the threshold increases, with no "correct" threshold at which to stop. **Such genuinely non-regular tail behaviour is, fortunately, rare in practice** — most physically, financially, or biologically motivated processes have tails that, even if complex, do not exhibit the oscillatory regular-variation failure of the constructed example above. However, the example serves as a useful reminder that the entire apparatus of Chapters 1-2 rests on an assumption — $F \in D(G_\xi)$ for some ξ — that, while extremely broad (encompassing essentially every named distribution in common statistical use), is not *universal*, and that persistent instability in threshold-selection diagnostics (Section 2.3), if not resolved by considering a wider range of thresholds or larger sample sizes, could in principle (though very rarely in practice) reflect a genuine failure of the domain-of-attraction assumption itself, rather than merely finite-sample noise.

Example — A practical near-example: mixtures with widely separated components

While the pure oscillatory pathology above is rarely encountered, a related and practically relevant phenomenon arises from *mixtures* with widely separated scales — e.g. F a mixture of 90% Exponential(rate 1) and 10% Exponential(rate 0.01, i.e. mean $100\times$ larger). For *moderate* thresholds u (say $u \in [1, 10]$), the mean-excess function $\hat{e}(u)$ is dominated by the first (light) component and behaves roughly as a constant ≈ 1 (Exponential mean-excess, Section 2.1). For *very high* thresholds ($u \gg 100$), the first component contributes negligibly to $\{X > u\}$, and $\hat{e}(u) \to 100$ (the second component's mean excess). **In between** (say $u \in [10, 100]$), $\hat{e}(u)$ transitions from ≈ 1 to ≈ 100 — a mean-excess plot exhibiting a pronounced "step" or rapid increase over this intermediate range, which could easily be *misread* as evidence of a strongly increasing linear trend (suggesting large $\xi > 0$, Section 2.1) when in fact the underlying mechanism is a finite mixture of two *exponential* (Gumbel-domain, $\xi = 0$) components — each individually well-behaved, but whose mixture exhibits transient behaviour over an intermediate threshold range that can masquerade as heavy-tailed (Fréchet-domain) behaviour if the threshold happens to fall within this transitional zone. This is an important practical caution when interpreting mean-excess plots (Section 2.3): **a mean-excess plot that appears to increase over the *available* range of thresholds does not, by itself, distinguish between "genuinely Fréchet-domain with this slope continuing indefinitely" and "Gumbel-domain mixture currently in a transitional regime that will eventually flatten at a higher threshold than is observable with the available data."** Substantive knowledge about the underlying data-generating mechanism — e.g. are there plausibly two or more distinct physical regimes or sub-populations contributing to the data? — is often essential to resolve this ambiguity, and is a recurring theme in the case studies of Appendix E (e.g. the tapered Gutenberg–Richter discussion of E.5, which is precisely a physically-motivated departure from a single-regime power law).

2.3 Threshold Selection: Mean Excess Plots and Hill Estimator

Overview

The PBdH theorem is asymptotic: it describes the limit $u \to x^*$. In practice, one must choose a *finite* threshold u, balancing the same bias-variance trade-off encountered for block size in Section 1.5: too low a threshold introduces bias (the GPD approximation is poor), too high a threshold leaves too few exceedances for stable estimation. This section develops the two principal graphical/numerical diagnostics for threshold choice: the mean excess plot and the Hill estimator.

The empirical mean excess plot. Given data $X_1, \dots, X_N$, define the *empirical mean excess function*

$$\hat{e}(u) = \frac{1}{|\{i : X_i > u\}|} \sum_{i : X_i > u} (X_i - u),$$

and plot $\hat{e}(u)$ against u for a range of candidate thresholds. By Section 2.1, if the true excess distribution above some threshold u_0 is exactly GPD(ξ, β), then for $u \geq u_0$,

$$e(u) = \frac{\beta + \xi(u - u_0)}{1 - \xi}$$

is *linear in* u. The threshold-selection rule is: choose the smallest u above which $\hat{e}(u)$ appears (within sampling variability) linear.

Example — Worked example: interpreting a mean excess plot

Suppose daily insurance claim sizes (in thousands of euros) yield the following empirical mean excesses at a sequence of thresholds:

u	$\hat{e}(u)$	Number of exceedances
10	8.2	950
20	14.1	410
30	19.8	180
40	25.9	75
50	38.4	22
60	71.0	6

From $u = 10$ to $u = 40$, $\hat{e}(u)$ increases roughly linearly with slope $\approx (25.9 - 8.2)/(40 - 10) = 0.59$, consistent with a GPD with $\xi/(1-\xi) \approx 0.59 \Rightarrow \xi \approx 0.37$. At $u = 50$ and especially $u = 60$, $\hat{e}(u)$ jumps sharply — but with only 22 and 6 exceedances respectively, this could be either (a) a genuine departure from linearity (e.g. the true tail is even heavier at extreme levels, a "tail of the tail" effect) or (b) pure sampling noise from a handful of observations. The standard error of $\hat{e}(u)$ scales roughly as $\hat{e}(u)/\sqrt{k}$ where k is the exceedance count, so at $u = 60$, SE $\approx 71/\sqrt{6} \approx 29$ — the apparent jump to 71.0 is well within one or two standard errors of the linear extrapolation ($\approx 0.59 \times 20 + 38.4 \approx 50.2$), and should not be over-interpreted. The recommended threshold here is $u \approx 30$–40: high enough for the GPD approximation to be reasonable, low enough to retain 75–180 exceedances for stable fitting.

The Hill estimator. For data already believed to be in the Fréchet domain ($\xi > 0$), the *Hill estimator* provides a direct, threshold-dependent estimate of $\alpha = 1/\xi$ (or equivalently of ξ), based on the order statistics. Given the k largest observations $X_{(N)} \geq X_{(N-1)} \geq \ldots \geq X_{(N-k+1)}$ (out of N total), the Hill estimator is

$$\hat{\xi}_{\text{Hill}} = \hat{\xi}_k = \frac{1}{k} \sum_{i=0}^{k-1} \log X_{(N-i)} - \log X_{(N-k)}.$$

Derivation of the Hill estimator

If $\overline{F}(x) = x^{-\alpha} L(x)$ with L slowly varying (the regular variation condition of Section 1.4), then for x large, $\log \overline{F}(x) \approx -\alpha \log x + \log L(x) \approx -\alpha \log x$ (the slowly-varying part contributes negligibly to the log on the relevant scale). This suggests $\log X$, conditional on $X > u$ for high u, is approximately Exponential with rate $\alpha = 1/\xi$ — i.e. $\alpha(\log X - \log u) \mid X > u \sim \text{Exp}(1)$ approximately. The Hill estimator is then simply the *method-of-moments / maximum-likelihood estimator of the mean of this exponential*, using the k largest observations with $u = X_{(N-k)}$ as the (random) threshold:

$$\hat{\xi}_k = \frac{1}{k} \sum_{i=0}^{k-1} \left[\log X_{(N-i)} - \log X_{(N-k)} \right],$$

the sample mean of the log-excesses above the k-th largest order statistic, on the log scale.

Asymptotic properties. Under appropriate conditions (in particular $k = k(N) \to \infty$ but $k/N \to 0$), the Hill estimator is consistent, $\hat{\xi}_k \to \xi$, and asymptotically normal,

$$\sqrt{k}(\hat{\xi}_k - \xi) \xrightarrow{d} \mathcal{N}(0, \xi^2),$$

provided $\xi > 0$ (the Hill estimator is specifically designed for, and only valid in, the Fréchet domain — applying it to Gumbel- or Weibull-domain data yields a meaningless or misleading estimate, often biased toward 0).

The Hill plot and its interpretation. Plotting $\hat{\xi}_k$ against k (or against the corresponding threshold $X_{(N-k)}$) produces the *Hill plot*. Ideally, $\hat{\xi}_k$ stabilizes (is roughly constant) over a range of k, and this stable value is taken as $\hat{\xi}$. In practice, Hill plots are notoriously difficult to read:

- For *small* k (very high threshold, few order statistics), $\hat{\xi}_k$ has high variance — the plot is "noisy" on the left.
- For *large* k (lower threshold, more order statistics), bias creeps in if the regular-variation approximation is imperfect at lower thresholds — the plot may show a systematic trend ("Hill horror plot") rather than a stable plateau.
- The presence of a slowly-varying component $L(x)$ that is not asymptotically constant (e.g. $L(x) = \log x$, common in some financial models) can produce a Hill plot with *no* stable region at any practically accessible sample size — the estimator remains biased for all k achievable with finite data.

Example — Numerical Hill plot illustration

Suppose $N = 2000$ i.i.d. observations from a Pareto distribution with true $\alpha = 2$ (so $\xi = 0.5$). The Hill estimator computed for various k:

k	$\hat{\xi}_k$	SE $= \hat{\xi}_k/\sqrt{k}$	95% CI
20	0.61	0.136	$(0.34, 0.88)$
50	0.54	0.076	$(0.39, 0.69)$
100	0.49	0.049	$(0.39, 0.59)$
200	0.52	0.037	$(0.45, 0.59)$
400	0.48	0.024	$(0.43, 0.53)$

All confidence intervals comfortably include the true value $\xi = 0.5$, and $\hat{\xi}_k$ fluctuates around 0.5 without strong trend — a "well-behaved" Hill plot, as expected for genuinely Pareto-distributed data. The width of the CI at $k = 20$ (about 0.54) versus $k = 400$ (about 0.10) illustrates the variance-reduction benefit of using more order statistics, *provided* bias from departure from exact regular variation does not dominate at the larger k — which, for genuinely Pareto data, it does not, by construction.

Threshold selection remains, to a significant degree, an art. Numerous automatic threshold-selection algorithms exist (minimizing asymptotic mean-squared error of the Hill estimator via bootstrap, e.g. the methods of Hall, or Drees–de Haan–Resnick; goodness-of-fit-based methods comparing GPD fits across candidate thresholds), but none is universally reliable. Best practice combines: (i) mean excess plots, (ii) Hill plots (if Fréchet domain is suspected), (iii) parameter stability plots (plotting $\hat{\xi}$ and $\hat{\beta}$ from GPD fits across a range of thresholds, looking for a stable region), and (iv) substantive domain knowledge about the physical, financial, or biological process generating the data. The sensitivity of downstream risk measures (Chapter 4) to threshold choice should always be explicitly reported, not hidden behind a single "optimal" threshold.

Automatic threshold selection: the Eye-Ball method formalized. Section 2.3 described

threshold selection as "to a significant degree, an art," relying on visual inspection of mean-excess and Hill plots. We now describe one widely-used *automated* procedure — not to replace visual inspection, but to provide a reproducible starting point and a cross-check.

Algorithm (Minimum AMSE threshold selection, Hall–Welsh type)

For a grid of candidate thresholds $u_1 < u_2 < \cdots < u_J$ (equivalently, candidate exceedance counts $k_1 > k_2 > \cdots > k_J$):

1. For each k_j, compute $\hat{\xi}_{k_j}$ (Hill estimator, Section 2.3, or GPD-MLE, Section 2.4) and an estimate $\hat{\rho}_{k_j}$ of the second-order parameter (Appendix C.1).
2. Estimate the asymptotic MSE: $\widehat{\mathrm{AMSE}}(k_j) = \frac{\hat{\xi}_{k_j}^2}{k_j} + \left[\hat{c} \cdot (n/k_j)^{\hat{\rho}_{k_j}}\right]^2$ for an estimated constant $\hat{c}$ (from the second-order fit).
3. Select $k^* = \arg\min_j \widehat{\mathrm{AMSE}}(k_j)$.

Example — Applying the algorithm to the running insurance example

Recall Section 2.3's mean-excess table, which suggested $u \in [30, 40]$ as reasonable. Suppose the automated AMSE-minimization procedure, applied to the same dataset across a finer grid $k \in \{50, 75, 100, 125, 150, 180, 215, 250\}$ (corresponding to thresholds roughly from $u \approx 45$ down to $u \approx 22$), produces:

k	$\hat{\xi}_k$	$\widehat{\mathrm{Var}} = \hat{\xi}_k^2/k$	$\widehat{\mathrm{Bias}}^2$	$\widehat{\mathrm{AMSE}}$
50	0.29	0.00168	0.00012	0.00180
75	0.31	0.00128	0.00031	0.00159
100	0.33	0.00109	0.00058	0.00167
125	0.34	0.00093	0.00095	0.00188
150	0.36	0.00086	0.00148	0.00234
180	0.34	0.00064	0.00210	0.00274

The minimum AMSE occurs at $k = 75$, suggesting a threshold corresponding to roughly $k = 75$ exceedances — **somewhat fewer exceedances than the $k = 180$ used in Section 2.4's worked example** (which corresponded to $u = 30$). At $k = 75$ (a higher threshold than $u = 30$), $\hat{\xi}_{75} = 0.31$, close to but slightly below the $\hat{\xi}_{180} = 0.34$ used previously. The AMSE-based recommendation would suggest the Section 2.4 analysis used a threshold slightly too low (too many exceedances, incurring more bias than necessary) — though the difference in $\hat{\xi}$ (0.31 vs 0.34) is well within one standard error of either estimate, illustrating that even a "principled" automated procedure, when applied to real data with realistic sample sizes, often yields recommendations that differ from a reasonable visual-inspection-based choice by an amount that is *statistically* immaterial, even if the AMSE values themselves vary noticeably across k. **The main value of automated procedures like this is less in identifying a single "correct" threshold** (no such single correct threshold exists, given the bias-variance trade-off is continuous, not sharply minimized) **than in providing a documented, reproducible, and bias-aware starting point** that can then be refined via the visual diagnostics (mean-excess plots, parameter stability plots) that remain, per Section 2.3's original discussion, an essential complement to any automated procedure.

Parameter stability plots. A complementary diagnostic to the Hill plot: for each candidate threshold u, plot $\hat{\xi}(u)$ and $\hat{\beta}(u) - \xi u$ (the latter being, by the threshold-stability relation of Section 2.1, expected to be approximately *constant* across u if the GPD model holds exactly

above some threshold — recall $\beta(u) = \beta_0 + \xi(u - u_0)$, so $\beta(u) - \xi u = \beta_0 - \xi u_0$, a constant). A parameter stability plot showing $\hat{\beta}(u) - \hat{\xi}(u) \cdot u$ approximately flat across a range of u provides direct visual evidence for the threshold-stability property itself (not merely for $\hat{\xi}(u)$ being stable), and is increasingly recommended as a primary diagnostic alongside (or in place of) the Hill plot, since it directly tests the GPD model's defining property rather than only one of its parameters.

2.4 Maximum Likelihood and Probability-Weighted Moment Estimation

Overview

Having selected a threshold u (Section 2.3) and obtained k exceedances $Y_1, \dots, Y_k$ (with $Y_i = X_i - u$ for $X_i > u$), we now estimate the GPD parameters (ξ, β). We develop two estimation methods in detail: maximum likelihood (ML), the most statistically efficient in large samples but occasionally numerically delicate; and the probability-weighted moments (PWM) method, more robust in small samples and for ξ near the boundary of the regular regime.

Maximum likelihood for the GPD. The log-likelihood for (ξ, β) given i.i.d. exceedances $Y_1, \dots, Y_k \sim H_{\xi,\beta}$, for $\xi \neq 0$, is

$$\ell(\xi, \beta) = -k \log \beta - \left(1 + \frac{1}{\xi}\right) \sum_{i=1}^{k} \log\left(1 + \xi \frac{Y_i}{\beta}\right),$$

valid provided $1 + \xi Y_i/\beta > 0$ for all i (automatically satisfied if $\xi \geq 0$; if $\xi < 0$, requires $Y_i < -\beta/\xi$ for all i, i.e. no exceedance exceeds the implied finite upper endpoint).

Score equations. Setting $\partial\ell/\partial\beta = 0$ and $\partial\ell/\partial\xi = 0$ yields a coupled nonlinear system; unlike many exponential-family models, there is no closed-form MLE for the GPD (except $\xi = 0$, where $\hat{\beta} = \bar{Y}$, the sample mean). Numerical optimization (Newton–Raphson, or profile likelihood over ξ with β concentrated out) is standard.

Asymptotic regime and regularity. The GPD MLE is asymptotically normal and efficient *for* $\xi > -1/2$:

$$\sqrt{k} \begin{pmatrix} \hat{\xi} - \xi \\ \hat{\beta}/\beta - 1 \end{pmatrix} \xrightarrow{d} \mathcal{N}\left(\mathbf{0}, (1+\xi) \begin{pmatrix} 1+\xi & -1 \\ -1 & 2 \end{pmatrix}\right).$$

For $\xi \leq -1/2$, the support of $H_{\xi,\beta}$ depends on the parameters in a way that violates standard regularity conditions (the Fisher information is not well-defined in the usual sense), and the MLE can have non-standard (even superefficient) asymptotic behaviour. As noted in Section 1.3 for the GEV, $\xi \leq -1/2$ corresponds to very short, sharply-bounded tails and is rarely the regime of interest in risk and reliability applications, but the boundary is worth knowing.

Example — Worked numerical example: GPD MLE for exceedance data

Continuing the insurance claims example of Section 2.3 with threshold $u = 30$ and $k = 180$ exceedances $Y_i = X_i - 30$. Suppose the sample mean and variance of the exceedances are $\bar{Y} = 15.2$ and $s_Y^2 = 580$.
A method-of-moments starting point: for the GPD, $\mathbb{E}[Y] = \beta/(1-\xi)$ and $\mathrm{Var}(Y) = \beta^2/[(1-\xi)^2(1-2\xi)]$ (for $\xi < 1/2$), so

$$\frac{\mathrm{Var}(Y)}{(\mathbb{E}[Y])^2} = \frac{1}{1-2\xi} \quad \Rightarrow \quad \hat{\xi}^{(0)} = \frac{1}{2}\left(1 - \frac{(\bar{Y})^2}{s_Y^2}\right) = \frac{1}{2}\left(1 - \frac{15.2^2}{580}\right) = \frac{1}{2}\left(1 - \frac{231.04}{580}\right) \approx \frac{1}{2}(1-0.3984) \approx 0.3$$

$$\hat{\beta}^{(0)} = \bar{Y}(1 - \hat{\xi}^{(0)}) \approx 15.2 \times 0.699 \approx 10.62.$$

Refining by Newton–Raphson on the score equations (details omitted) might converge to, say, $\hat{\xi}_{\text{ML}} = 0.34$ (SE 0.094), $\hat{\beta}_{\text{ML}} = 9.8$ (SE 1.6), with $\widehat{\text{Cov}}(\hat{\xi}, \hat{\beta}) \approx -0.11$ — the negative correlation between $\hat{\xi}$ and $\hat{\beta}$ is typical and reflects the GPD's parameter redundancy at the boundary (large ξ with small β can mimic small ξ with large β over the bulk of the exceedance range, with differences emerging only in the extreme tail of the exceedance distribution itself).

Probability-weighted moments (PWM). The PWM method, developed by Hosking and Wallis (1987) specifically for the GPD and GEV, is based on moments of the form

$$M_{1,0,s} = \mathbb{E}\left[Y\,(1 - H(Y))^s\right], \qquad s = 0, 1, 2, \ldots$$

For the GPD (valid for $\xi < 1$):

$$M_{1,0,0} = \mathbb{E}[Y] = \frac{\beta}{1-\xi}, \qquad M_{1,0,1} = \mathbb{E}[Y(1 - H(Y))] = \frac{\beta}{(1-\xi)(2-\xi)}.$$

Equating to sample estimates $\hat{M}_{1,0,0} = \bar{Y}$ and $\hat{M}_{1,0,1} = \frac{1}{k}\sum_i \left(1 - \frac{i-0.35}{k}\right) Y_{(i)}$ (using a standard plotting-position estimator on the order statistics $Y_{(1)} \le \ldots \le Y_{(k)}$), and solving the resulting two equations:

$$\frac{M_{1,0,0}}{M_{1,0,0} - 2M_{1,0,1}} = \frac{1}{2-\xi} \cdot \frac{2-\xi}{1} \quad \Rightarrow \quad \hat{\xi}_{\text{PWM}} = \frac{\bar{Y}}{\bar{Y} - 2\hat{M}_{1,0,1}} - 2, \qquad \hat{\beta}_{\text{PWM}} = (1 - \hat{\xi}_{\text{PWM}})\bar{Y}.$$

Comparing ML and PWM. Simulation studies (Hosking & Wallis 1987 and many subsequent works) consistently find:

- For $\xi \in (-0.5, 0.5)$ and moderate-to-large samples ($k \gtrsim 50$), ML and PWM give similar results, with ML having slightly lower asymptotic variance.
- For ξ close to or above 0.5, PWM estimators of β become unstable (since $M_{1,0,1}$ requires $\xi < 1$ to exist, and higher PWMs used in some variants require even smaller ξ), while ML remains well-defined for $\xi < 1$ (existence of likelihood) though with poor finite-sample behaviour.
- For small samples ($k \lesssim 50$), PWM is often *less biased* for ξ, though with comparable or slightly larger variance — PWM is frequently preferred in hydrological applications where sample sizes (decades of annual data) are inherently small.
- Neither method handles $\xi \ge 1$ well: ML likelihood is unbounded as $\xi \to \infty$ along certain paths (a known pathology of the GPD/GEV likelihood for extreme shape values), requiring constrained optimization or penalized likelihood in practice.

A Bayesian alternative. Both ML and PWM produce *point estimates* with associated (asymptotic, possibly poorly-calibrated for small k) standard errors. An increasingly standard alternative — especially for risk applications where uncertainty in ξ propagates directly into uncertainty in extreme quantiles (Section 2.5) and risk measures (Chapter 4) — is full Bayesian inference for (ξ, β), using e.g. weakly informative priors (a common choice constrains ξ to a plausible range like $(-1, 1)$ via a suitable prior, reflecting that $|\xi| > 1$ is physically implausible for most applications) and Markov

Chain Monte Carlo (MCMC) to obtain the full posterior. We return to Bayesian extreme value inference in depth in Chapter 4 of this volume (in the risk-measure context) and systematically in Part V.

The relationship to Section 1.5's block-maxima estimation. Both the GEV (block maxima) and GPD (threshold exceedances) approaches estimate the *same* underlying ξ (by PBdH, Section 2.2), but typically with different effective sample sizes for a given amount of raw data: m block maxima versus k threshold exceedances, where k can usually be made considerably larger than m for the same total data (since POT uses *all* sufficiently large observations, not one per block). This is the principal practical advantage of POT over block maxima, at the cost of the additional threshold-selection step (Section 2.3) and the need to model the exceedance *times* (typically via a Poisson process, see the chapter's closing remarks) in addition to the exceedance *magnitudes*.

The L-moments approach: a robust alternative to PWM. Closely related to the probability-weighted moments of Section 2.4, but with a slightly different (and in some respects more transparent) formulation, are the *L-moments*, defined via expectations of linear combinations of order statistics:

$$\lambda_1 = \mathbb{E}[X_{(1:1)}], \qquad \lambda_2 = \tfrac{1}{2}\mathbb{E}[X_{(2:2)} - X_{(1:2)}], \qquad \lambda_3 = \tfrac{1}{3}\mathbb{E}[X_{(3:3)} - 2X_{(2:3)} + X_{(1:3)}],$$

where $X_{(i:n)}$ denotes the i-th order statistic of a sample of size n. λ_1 is the mean; λ_2 is a measure of scale (related to, but distinct from, the standard deviation — it is half the expected difference between two independent draws' larger and smaller values); λ_3 relates to skewness via the *L-skewness* $\tau_3 = \lambda_3/\lambda_2$.

L-moments of the GPD

For the GPD with $\xi < 1$:

$$\lambda_1 = \frac{\beta}{1-\xi}, \qquad \lambda_2 = \frac{\beta}{(1-\xi)(2-\xi)}, \qquad \tau_3 = \frac{1-\xi}{3-\xi}.$$

Why L-moments? Unlike ordinary moments (Section 3.1's discussion of moment existence), L-moments exist as long as $\mathbb{E}[X] < \infty$ — $\lambda_2, \lambda_3, \ldots$ exist whenever λ_1 does, since they are linear combinations of order statistic expectations rather than involving higher powers of X. This means **L-moment estimators remain well-defined (with finite asymptotic variance) for a wider range of ξ than ordinary-moment-based PWM estimators**, which (as Section 2.4 noted) become unstable as $\xi \to 1$ (where $M_{1,0,1}$ itself approaches a boundary of existence). For $\xi \in (0, 1)$, L-moment estimators of ξ have asymptotic variance that remains finite throughout this range, including as $\xi \to 1$, whereas PWM-based estimators' variance can blow up.

Example — Numerical comparison: PWM vs. L-moments for ξ near 1

Suppose true $\xi = 0.85$ (a very heavy tail, mean barely finite). Simulating $k = 100$ GPD exceedances and comparing estimators:
PWM: the formula $\hat{\xi}_{\text{PWM}} = \bar{Y}/(\bar{Y} - 2\hat{M}_{1,0,1}) - 2$ (Section 2.4) involves $\hat{M}_{1,0,1}$, whose population value $M_{1,0,1} = \beta/[(1-\xi)(2-\xi)]$ at $\xi = 0.85$ equals $\beta/(0.15 \times 1.15) = \beta/0.1725 \approx 5.80\beta$ — a large value relative to β, and the *sample* $\hat{M}_{1,0,1}$ (involving products of order statistics with weights $(1-(i-0.35)/k)$) has high variance when a few large order statistics dominate the sum (again, the single-big-jump principle, Section 3.3, operating within the estimator itself). Simulated SD($\hat{\xi}_{\text{PWM}}$) across replications

≈ 0.31.
L-moments: $\tau_3 = (1-\xi)/(3-\xi) = (1-0.85)/(3-0.85) = 0.15/2.15 \approx 0.0698$. The L-skewness estimator $\hat{\tau}_3 = \hat{\lambda}_3/\hat{\lambda}_2$ involves only *linear* combinations of order statistics (no products), giving it more favourable variance properties for heavy tails. Simulated $\mathrm{SD}(\hat{\xi}_{\mathrm{Lmom}}) \approx 0.22$ — roughly 30% smaller than PWM's 0.31.
For ξ in the more "moderate" range explored in Section 2.4's main worked example ($\xi \approx 0.34$), PWM and L-moments give very similar performance (both SD ≈ 0.09–0.10, comparable to the ML standard error of 0.094 reported there) — the L-moment advantage emerges specifically as ξ approaches and exceeds 0.5–1, precisely the regime — as Appendix G's operational risk case study illustrated — where accurate tail-index estimation matters most and is hardest to achieve. This makes L-moments a recommended addition to the analyst's toolkit specifically as a robustness check whenever $\hat{\xi}_{\mathrm{ML}}$ or $\hat{\xi}_{\mathrm{PWM}}$ exceeds approximately 0.5.

2.5 Return Levels and Return Periods

Overview

We now connect the fitted GPD (or GEV) model back to the practical quantities that motivate extreme value analysis in engineering, hydrology, and risk management: *return levels* (a level expected to be exceeded, on average, once every T years/observations) and *return periods* (the expected waiting time between exceedances of a given level). We derive these from the POT model, propagate parameter uncertainty via the delta method, and discuss the subtle non-stationary extensions relevant to a changing climate or evolving risk environment.

Return level from the POT model. Suppose exceedances of threshold u occur, on average, λ times per year (estimated as $\hat{\lambda} = k/(\text{years of data})$), and excess magnitudes follow $\mathrm{GPD}(\xi, \beta)$. The annual probability that the level $x > u$ is exceeded *at least once* is

$$p_{\mathrm{exc}}(x) = 1 - \left[1 - \overline{H}_{\xi,\beta}(x-u)\right]^{\lambda} \approx \lambda\, \overline{H}_{\xi,\beta}(x-u)$$

for small $\lambda \overline{H}_{\xi,\beta}(x-u)$ (Poisson approximation: the number of exceedances of u in a year is approximately $\mathrm{Poisson}(\lambda)$, and each independently exceeds x with probability $\overline{H}_{\xi,\beta}(x-u)$, so the annual count exceeding x is approximately $\mathrm{Poisson}(\lambda \overline{H}_{\xi,\beta}(x-u))$, and $P(\text{count} \geq 1) \approx$ its mean for small means).

Definition (Return level)

The *T-year return level* x_T is the level whose annual exceedance probability is $1/T$:

$$\lambda\, \overline{H}_{\xi,\beta}(x_T - u) = \frac{1}{T} \quad \Longleftrightarrow \quad \overline{H}_{\xi,\beta}(x_T - u) = \frac{1}{\lambda T}.$$

Solving for $\xi \neq 0$:

$$\left(1 + \xi \frac{x_T - u}{\beta}\right)^{-1/\xi} = \frac{1}{\lambda T} \quad \Rightarrow \quad x_T = u + \frac{\beta}{\xi}\left[(\lambda T)^{\xi} - 1\right].$$

For $\xi = 0$: $x_T = u + \beta \log(\lambda T)$.

Interpretation. x_T is *not* "the maximum level that will occur in T years" (a common misreading); it is the level with annual exceedance probability $1/T$, equivalently the level whose *expected* number of exceedances over T years is 1. The actual maximum over T years is itself a random variable whose distribution can be derived from the GEV (Section 1.3) fitted to T-year block maxima — and indeed, the return-level formula above is mathematically consistent with (an asymptotic approximation to) the GEV quantile formula of Section 1.3, with the GEV parameters expressed in terms of the POT parameters (λ, ξ, β, u) via $\mu_{\text{GEV}} = u + \frac{\beta}{\xi}(\lambda^\xi - 1)$ and $\sigma_{\text{GEV}} = \beta\lambda^\xi$.

Example — Worked example: 50- and 100-year return levels with uncertainty

Continuing the insurance/hydrological example: threshold $u = 30$, $\hat{\xi} = 0.34$, $\hat{\beta} = 9.8$, and suppose exceedances of $u = 30$ occur on average $\hat{\lambda} = 4.5$ times per year (e.g. daily data with 180 exceedances over 40 years).
50-year return level:

$$\lambda T = 4.5\times 50 = 225, \qquad (225)^{0.34} = \exp(0.34\times \ln 225) = \exp(0.34\times 5.416) = \exp(1.841) \approx 6.30.$$

$$x_{50} = 30 + \frac{9.8}{0.34}(6.30 - 1) = 30 + 28.82 \times 5.30 \approx 30 + 152.7 \approx 182.7.$$

100-year return level:

$$\lambda T = 450, \qquad (450)^{0.34} = \exp(0.34 \times 6.109) = \exp(2.077) \approx 7.98.$$

$$x_{100} = 30 + 28.82 \times (7.98 - 1) \approx 30 + 201.1 \approx 231.1.$$

Uncertainty via the delta method. Write $x_T = g(\xi, \beta; u, \lambda, T)$. The delta method gives

$$\text{Var}(\hat{x}_T) \approx \left(\frac{\partial g}{\partial \xi}\right)^2 \text{Var}(\hat{\xi}) + \left(\frac{\partial g}{\partial \beta}\right)^2 \text{Var}(\hat{\beta}) + 2\frac{\partial g}{\partial \xi}\frac{\partial g}{\partial \beta}\text{Cov}(\hat{\xi}, \hat{\beta}).$$

With $\partial g/\partial \beta = \frac{1}{\xi}[(\lambda T)^\xi - 1]$ and $\partial g/\partial \xi = -\frac{\beta}{\xi^2}[(\lambda T)^\xi - 1] + \frac{\beta}{\xi}(\lambda T)^\xi \log(\lambda T)$, and using the earlier estimates $\text{Var}(\hat{\xi}) \approx 0.094^2 = 0.0088$, $\text{Var}(\hat{\beta}) \approx 1.6^2 = 2.56$, $\text{Cov}(\hat{\xi}, \hat{\beta}) \approx -0.11$:
For $T = 100$: $\partial g/\partial \beta = \frac{1}{0.34}\times 6.98 \approx 20.5$; $\partial g/\partial \xi = -\frac{9.8}{0.34^2}\times 6.98 + \frac{9.8}{0.34}\times 7.98 \times \log 450 \approx -591.8 + 28.82 \times 7.98 \times 6.109 \approx -591.8 + 1404.7 \approx 812.9$.

$$\text{Var}(\hat{x}_{100}) \approx 812.9^2\times 0.0088 + 20.5^2 \times 2.56 + 2\times 812.9 \times 20.5 \times (-0.11) \approx 5816 + 1076 - 3666 \approx 3226.$$

$$\text{SE}(\hat{x}_{100}) \approx \sqrt{3226} \approx 56.8.$$

A 95% confidence interval for the 100-year return level is approximately $231.1 \pm 1.96 \times 56.8 \approx (119.8, 342.4)$ — a remarkably wide interval, spanning nearly a factor of three, driven primarily by the large $\text{Var}(\hat{\xi})$ term and its (here, partially offsetting but still net positive) covariance structure with $\hat{\beta}$. This enormous uncertainty band is *typical* of extreme value extrapolation and is a central, sobering message of this entire chapter: **point estimates of return levels at long return periods convey a false sense of precision; the associated uncertainty intervals, when correctly propagated, are often embarrassingly wide.**

Non-stationary extensions. The formulas above assume (ξ, β, λ) are constant over time — a *stationarity* assumption increasingly questioned in climate-affected hydrological applications and in financial/economic risk settings with structural breaks. Non-stationary exten-

sions allow parameters to depend on covariates (e.g. time, CO_2 concentration, an economic regime indicator):

$$\mu_t = \mu_0 + \mu_1 t, \quad \text{or} \quad \log \beta_t = \beta_0 + \beta_1 t,$$

fitted by maximum likelihood with time-varying parameters. The notion of "return period" itself becomes ambiguous under non-stationarity — several competing definitions exist (the *expected waiting time* definition and the *effective return level for a fixed future period* definition can give materially different numbers), a subtlety that has caused considerable confusion in the climate-risk literature and that we revisit when discussing non-stationary risk in Chapter 4.

Return levels and coherent risk measures: a preview. The return-level concept is closely related to, but not identical with, the *Value-at-Risk* (VaR) concept central to Chapter 4: $\text{VaR}_{1-1/T}$ for an annual loss distribution is essentially the T-year return level for that loss. The extreme value machinery of this chapter — GEV/GPD fitting, threshold selection, uncertainty propagation — is therefore not a side topic but the *primary toolkit* for estimating VaR and CVaR at the extreme quantiles (e.g. 99.9% or 99.97%, corresponding to $T = 1000$ or $T \approx 3333$) demanded by banking and insurance regulation, where empirical data essentially never directly populates the quantile of interest and extrapolation via (ξ, β) is unavoidable.

PART 3

Heavy Tails, Power Laws and Regular Variation

The mathematics of regular variation underlying power-law phenomena across physical, biological, and social systems.

Confidence intervals via profile likelihood for return levels. Section 2.5's delta-method confidence interval for x_{100} relied on asymptotic normality of $(\hat{\xi}, \hat{\beta})$ and a first-order Taylor expansion of $x_T = g(\xi, \beta)$. For ξ not small (as in the worked example, $\hat{\xi} = 0.34$), the delta method can be inaccurate — the true sampling distribution of $\hat{x}_T$ is often noticeably *right-skewed* (since x_T depends on ξ through $(\lambda T)^\xi$, an increasing *convex* function of ξ for $\lambda T > 1$, so a symmetric distribution for $\hat{\xi}$ translates into an asymmetric one for $\hat{x}_T$ — the same convexity phenomenon driving the Bayesian results of Appendices D.1 and K.3, now appearing in a purely frequentist context).

Profile likelihood interval for x_T

Reparametrize the GPD likelihood (Section 2.4) in terms of (x_T, ξ) instead of (β, ξ), using $\beta = \frac{\xi(x_T - u)}{(\lambda T)^\xi - 1}$ (inverting Section 2.5's return-level formula). The profile log-likelihood for x_T is

$$\ell_p(x_T) = \max_{\xi} \ell(\xi, \beta(x_T, \xi)).$$

A $100(1-\alpha)\%$ profile likelihood confidence interval for x_T is $\{x_T : 2[\ell(\hat{\xi}, \hat{\beta}) - \ell_p(x_T)] \leq \chi^2_{1,1-\alpha}\}$, exactly analogous to the profile likelihood interval for ξ alone (Section 1.5's closing discussion).

Example — Comparing delta-method and profile-likelihood intervals

For the running example ($\hat{\xi} = 0.34, \hat{\beta} = 9.8, u = 30, \lambda = 4.5, T = 100$, giving $\hat{x}_{100} \approx 231.1$ and delta-method SE ≈ 56.8, Section 2.5):

The delta-method 95% CI is symmetric: $231.1 \pm 1.96 \times 56.8 \approx (119.8, 342.4)$.
A profile-likelihood CI, by contrast, is typically *asymmetric*, reflecting the right-skew noted above. Suppose numerical profiling yields $(138.2, 378.6)$ — the lower bound is *higher* than the delta method's $(138.2$ vs. $119.8)$ while the upper bound is *higher still* $(378.6$ vs. $342.4)$: **the profile-likelihood interval is shifted and stretched toward larger values relative to the symmetric delta-method interval**, consistent with the convexity-driven right-skew. The profile interval's width $(378.6 - 138.2 = 240.4)$ is comparable to the delta method's $(342.4 - 119.8 = 222.6)$, but its *location* differs — the profile interval suggests less probability mass at very low return levels (below ≈ 138) than the delta method does, and correspondingly more mass at very high return levels (above ≈ 342).
Practical implication. If a regulatory or engineering threshold is set at, say, the lower end of a confidence interval (a conservative choice from one perspective — using a lower bound on the "true" return level might seem to understate required capacity, but is sometimes used when the relevant concern is *not under-designing*, i.e. when a too-low estimate of x_T would lead to insufficient capacity, so a *lower* confidence bound on x_T that is still reasonably high provides some assurance against x_T being even lower than thought — though, more commonly, *upper* bounds are used for conservative design, as in Appendix E.1's discussion), the choice between delta-method (119.8) and profile-likelihood (138.2) lower bounds differs by about 18.4 units, or roughly 15% of the lower bound itself — a non-trivial difference for design purposes, arising purely from the choice of *interval construction method*, holding the data and the GPD model itself fixed. This is yet another instance — alongside threshold choice (Section 2.3, Appendix G), estimation method (Section 2.4, this appendix's L-moments discussion), copula family (Appendix J), prior specification (Appendix K), and non-stationarity definitional choices (Appendix I) — of a seemingly secondary methodological choice materially affecting the final reported risk quantity, reinforcing this volume's central message about the pervasiveness of such sensitivities in extreme value-based risk analysis.

3.1 Regularly Varying Functions and Karamata's Theorem

Overview

Regular variation is the precise mathematical language for "power-law-like behaviour at infinity" that we have used informally since Section 1.4. This section develops the theory rigorously: slowly varying functions, the characterization theorem of Karamata, and Karamata's theorem on integrals of regularly varying functions — the technical foundation underlying every heavy-tail result in this volume, from the Hill estimator to the principle of a single big jump (Section 3.3).

Definition (Slowly and regularly varying functions)

A positive, measurable function $L : (0, \infty) \to (0, \infty)$ is *slowly varying at infinity* (written $L \in RV_0$) if

$$\lim_{x \to \infty} \frac{L(tx)}{L(x)} = 1 \qquad \text{for all } t > 0.$$

A positive measurable function f is *regularly varying with index* $\rho \in \mathbb{R}$ (written $f \in$

RV_ρ) if

$$f(x) = x^\rho L(x)$$

for some $L \in RV_0$; equivalently, $\lim_{x\to\infty} f(tx)/f(x) = t^\rho$ for all $t > 0$.

Examples. $L(x) = \log x$, $L(x) = \log\log x$, $L(x) = (\log x)^\beta$ for any $\beta \in \mathbb{R}$, and $L(x) \to c$ (any positive constant) are all slowly varying. Note that $L(x) = x^\epsilon$ for $\epsilon \neq 0$ is *not* slowly varying — slow variation permits much slower-than-any-power growth or decay, but not even an arbitrarily small power. $f(x) = x^2 \log x \in RV_2$; $f(x) = x^{-\alpha}(\log x)^{-\gamma} \in RV_{-\alpha}$ for any γ; $f(x) = e^x \notin RV_\rho$ for any ρ (exponential functions are not regularly varying — this is precisely why the Gumbel domain, Section 1.4, is *not* characterized by regular variation of $\overline{F}$, but rather by the von Mises condition).

Theorem (Karamata's characterization / representation theorem)

$L \in RV_0$ if and only if L can be written as

$$L(x) = c(x) \exp\left(\int_{x_0}^{x} \frac{\epsilon(t)}{t}\, dt \right)$$

for $x \geq x_0$, where $c(x) \to c \in (0, \infty)$ and $\epsilon(t) \to 0$ as $t, x \to \infty$.

This representation makes precise the intuition that slowly varying functions are "asymptotically constant on a logarithmic scale" — the integral $\int \epsilon(t)/t\, dt$ grows slower than any positive power of x because $\epsilon(t) \to 0$.

Theorem (Karamata's theorem on integrals)

If $f \in RV_\rho$ with $\rho > -1$, then

$$\int_0^x f(t)\, dt \sim \frac{x f(x)}{\rho + 1} \qquad \text{as } x \to \infty,$$

i.e. $\int_0^x f \in RV_{\rho+1}$. If $f \in RV_\rho$ with $\rho < -1$, then $\int_x^\infty f(t)\, dt$ converges and

$$\int_x^\infty f(t)\, dt \sim \frac{-x f(x)}{\rho + 1} = \frac{x f(x)}{|\rho + 1|} \qquad \text{as } x \to \infty,$$

i.e. $\int_x^\infty f \in RV_{\rho+1}$ (note $\rho + 1 < 0$). The boundary case $\rho = -1$ requires separate treatment (the integral may or may not converge depending on the slowly varying factor; e.g. $\int^\infty dx/(x \log x)$ diverges while $\int^\infty dx/(x(\log x)^2)$ converges).

Example — Application: tail of a sum of regularly varying terms, and moment existence

Let X have density $f(x) \sim Cx^{-\alpha-1}$ as $x \to \infty$ (so $f \in RV_{-\alpha-1}$), corresponding to $\overline{F}(x) \sim \frac{C}{\alpha} x^{-\alpha} \in RV_{-\alpha}$ by Karamata's integral theorem (since $\overline{F}(x) = \int_x^\infty f(t)dt$ and $-\alpha - 1 < -1$).

Existence of moments. $\mathbb{E}[X^k] = \int_0^\infty k x^{k-1} \overline{F}(x)\, dx$ (for $X \geq 0$). The integrand is $\sim kC' x^{k-1-\alpha}$, which is in $RV_{k-1-\alpha}$. By Karamata, this integral converges (at ∞) iff $k - 1 - \alpha < -1$, i.e. $k < \alpha$. This recovers, via regular variation, the familiar fact that a Pareto-tailed distribution with index α has finite moments only up to (but not including)

order α — exactly the moment-existence table of Section 1.3 for the GEV/Fréchet case, now derived from first principles via Karamata's theorem rather than asserted.
Numerical check. For $\alpha = 2.5$ (a typical value for, e.g., city-size or firm-size distributions): moments of order 1 and 2 (mean and variance) are finite, but the third moment ($k = 3 > 2.5$) is infinite. The skewness and kurtosis of such data, computed from a finite sample, will appear to converge to *some* value as sample size grows, but this apparent convergence is illusory — the sample skewness/kurtosis statistics do not converge to a finite population value but instead grow (in expectation and in probability) without bound, a phenomenon sometimes called the "preasymptotic" behaviour of heavy-tailed sample moments, and a frequent source of spurious findings in empirical studies that report higher moments of heavy-tailed financial or network data without checking α.

The uniform convergence theorem. A technically crucial result (which we state without proof) is that for $f \in RV_\rho$, the convergence $f(tx)/f(x) \to t^\rho$ is *uniform* for t in any compact subset of $(0, \infty)$. This uniformity is what permits the interchange of limits required in proofs such as the PBdH theorem (Section 2.2) and the convergence of the Hill estimator (Section 2.3) — pointwise convergence of the defining ratio alone would not suffice for these results.

Regular variation as the unifying language. Every major result in Chapters 1–2 can be restated in terms of regular variation: $F \in D(\Phi_\alpha)$ (Fréchet domain) iff $\overline{F} \in RV_{-\alpha}$; the GPD scale function $\beta(u)$ in PBdH is determined by the slowly varying part of $\overline{F}$; the Hill estimator's consistency relies on the uniform convergence theorem applied to $\overline{F}$. Chapter 3 develops regular variation as a tool in its own right because its applications extend far beyond extreme value theory proper — to the tail behaviour of sums (Section 3.3), to multivariate dependence structures (Section 3.5), and to the power-law phenomena of Section 3.2, which arise across economics, physics, biology, and network science.

Potter's theorem and uniform bounds. The uniform convergence theorem cited in Section 3.1 has a quantitative counterpart — *Potter's theorem* — that is indispensable for proving convergence results in extreme value theory at a level of rigor beyond what pure pointwise limits provide.

Theorem (Potter's theorem)

If $L \in RV_0$, then for every $\epsilon > 0$ there exists $x_0 = x_0(\epsilon)$ such that for all $x, y \geq x_0$:

$$(1-\epsilon)\min\left(\frac{x}{y}, \frac{y}{x}\right)^\epsilon \leq \frac{L(x)}{L(y)} \leq (1+\epsilon)\max\left(\frac{x}{y}, \frac{y}{x}\right)^\epsilon.$$

In words: for large enough x, y, the ratio $L(x)/L(y)$ is within an ϵ-power-law band around 1 — the slowly varying function cannot fluctuate faster than any power law, but can still fluctuate within that bound (which is what makes, e.g., $L(x) = \log x$ slowly varying: $\log(tx)/\log x \to 1$ for every t, but the approach can be slow, and Potter's theorem quantifies the rate of approach uniformly in t).

Example — Potter's theorem applied to the Hill estimator's consistency proof

The Hill estimator $\hat{\xi}_k = \frac{1}{k}\sum_{i=0}^{k-1}[\log X_{(N-i)} - \log X_{(N-k)}]$ is consistent for $\xi = 1/\alpha$ (Section 2.3). A key step in the proof involves showing that $X_{(N-i)}/X_{(N-k)}$ concentrates

around $(k/i)^{\xi}$ for $i = 0, \dots, k-1$. This uses the fact that, for $F \in RV_{-\alpha}$, $\overline{F}(x) = x^{-\alpha}L(x)$ with L slowly varying, and the ratio

$$\frac{P(X > tx)}{P(X > x)} = \frac{(tx)^{-\alpha}L(tx)}{x^{-\alpha}L(x)} = t^{-\alpha}\frac{L(tx)}{L(x)}.$$

For L slowly varying, $L(tx)/L(x) \to 1$ pointwise, but Potter's theorem guarantees this ratio is bounded between $(1 \pm \epsilon)(t^{-\alpha} \cdot t^{\pm\epsilon})$ uniformly for $x \geq x_0$ — this *uniform* control is what allows the pointwise convergence of $L(tx)/L(x) \to 1$ to be upgraded to the uniform convergence of the order-statistic ratios needed for the consistency proof. Without Potter's theorem, establishing consistency via pointwise limits alone would require additional technical steps that clutter the core argument; with it, the proof is essentially: apply Potter's theorem, obtain uniform bounds, invoke an appropriate law of large numbers for the (dependent) log-spacings of order statistics.

Integration and summation theorems: a richer picture. Karamata's theorem (Section 3.1) gives the first-order asymptotic of integrals of regularly varying functions. For many applications — including the derivation of return-level formulas at second order, and the bias analysis of Appendix C.1 — one needs the next term. The full asymptotic expansion, under second-order regular variation (Appendix C.1), is

$$\int_x^{\infty} f(t)\,dt = \frac{xf(x)}{|\rho+1|}\left[1 + \frac{A(x)}{1+\rho} + o(A(x))\right],$$

where $A(x)$ is the second-order auxiliary function of Appendix C.1. This refined formula is the basis of the bias-corrected Hill estimator's derivation (Appendix C.1): the $O(A(x))$ correction term, when applied to the asymptotic expansion of $\hat{\xi}_k$'s expectation, gives the (first-order) bias $\approx c \cdot A(n/k)$ and thereby the MSE-optimal k that balances this bias against the variance ξ^2/k.

3.2 Power-Law Distributions: Definition, Estimation, and Pitfalls

Overview

Power-law distributions — $P(X > x) \sim Cx^{-\alpha}$ — are ubiquitous in empirical claims across many fields: city sizes, word frequencies, earthquake magnitudes, firm sizes, web link counts, and financial losses have all been described as "power-law distributed." This section provides the precise definitions, the correct (and frequently misused) estimation methodology, and a frank discussion of the pitfalls — both statistical and conceptual — that have led to substantial controversy and over-claiming in the power-law literature.

Definition and equivalent formulations. A random variable $X \geq x_{\min} > 0$ follows a power law (Pareto distribution) with tail index $\alpha > 0$ if

$$P(X > x) = \left(\frac{x}{x_{\min}}\right)^{-\alpha}, \qquad x \geq x_{\min}.$$

The density is $f(x) = \alpha x_{\min}^{\alpha} x^{-\alpha-1}$. Equivalently, $\log X - \log x_{\min}$, conditional on $X \geq x_{\min}$, is Exponential(α) — the basis of the Hill estimator (Section 2.3) and of the standard "log-log plot" diagnostic.

The log-log plot and its dangers. A complementary CDF $\overline{F}(x) = Cx^{-\alpha}$ satisfies $\log \overline{F}(x) = \log C - \alpha \log x$ — a straight line on log-log axes, with slope $-\alpha$. The (in)famous practice of "eyeballing" a log-log plot of empirical $\overline{F}$ and fitting a line by ordinary least squares (OLS) is now widely recognized as statistically unsound, for several compounding reasons:

- OLS on log-log axes implicitly assumes additive Gaussian errors on the log scale, which is not the correct error structure for order-statistic-based empirical CDFs (errors are correlated and heteroskedastic, growing in the tail).
- The choice of $x_{\min}$ (where the power law is assumed to begin) dramatically affects the fitted α, and OLS provides no principled way to choose it.
- Binning of the data (for histogram-based log-log plots) introduces additional, often severe, biases — particularly with logarithmic binning, where bin width itself scales with x.

Recommended methodology (Clauset–Shalizi–Newman, 2009)

Step 1. Estimate α via maximum likelihood (for continuous data, this reduces to the Hill estimator of Section 2.3, restricted to $X \geq x_{\min}$):

$$\hat{\alpha} = 1 + n \left[\sum_{i=1}^{n} \log \frac{X_i}{x_{\min}} \right]^{-1}.$$

Step 2. Estimate $x_{\min}$ by minimizing the Kolmogorov–Smirnov (KS) distance between the empirical CDF (restricted to $X \geq x_{\min}$) and the fitted Pareto CDF, over candidate values of $x_{\min}$.
Step 3. Assess *goodness of fit* via a bootstrap/Monte Carlo test: generate many synthetic datasets from the fitted power law (with the same n and $x_{\min}$), compute the KS distance for each, and compare to the KS distance of the original data. A small p-value (fraction of synthetic KS distances exceeding the observed one) indicates the power law is implausible.
Step 4. Compare to alternative distributions (log-normal, exponential, stretched exponential, power law with exponential cutoff) via likelihood ratio tests (Vuong's test), since many of these alternatives can mimic a power law over a finite range.

Example — Worked example: why the choice of $x_{\min}$ matters

Suppose data are generated from a distribution that is exactly Pareto($\alpha = 2.5$, $x_{\min} = 1$) for $x \geq 1$, but with an additional excess of small values below $x = 1$ (a common feature of real data: power laws typically hold only in the tail, with the bulk following a different distribution).
If one (incorrectly) sets $x_{\min} = 0.1$ and applies the Hill-type MLE $\hat{\alpha} = 1 + n\left[\sum \log(X_i/0.1)\right]^{-1}$, the extra mass below $x = 1$ contributes large $\log(X_i/0.1)$ terms (since $X_i/0.1$ can be as small as 1, giving $\log(X_i/0.1) \approx 0$ for $X_i \approx 0.1$, but the *distribution* of these small values, if not itself Pareto with $\alpha = 2.5$, biases the sum). A typical outcome is $\hat{\alpha}_{x_{\min}=0.1} \approx 1.7$ — substantially *lower* than the true $\alpha = 2.5$, i.e. the estimated tail appears *heavier* than it truly is, because the non-power-law bulk has been incorrectly included.
With the correct (KS-minimizing) $x_{\min} \approx 1$, $\hat{\alpha}_{x_{\min}=1} \approx 2.48$ — close to the true value. This illustrates the single most common error in applied power-law fitting: **including**

data below the true onset of power-law behaviour systematically biases $\hat{\alpha}$ toward smaller (heavier-tail) values, potentially leading to dramatic over-statement of tail risk.

The "fat middle" problem and alternatives to pure power laws. Many empirical distributions exhibit power-law-like behaviour in the extreme tail but a different shape (log-normal, Weibull, or exponentially-truncated power law) over the bulk and even the moderate tail. The *power law with exponential cutoff*,

$$\overline{F}(x) \propto x^{-\alpha} e^{-\lambda x},$$

nests both pure power law ($\lambda = 0$) and exponential ($\alpha = 0$) as special cases, and is often a better-fitting and more physically motivated model for finite-size or finite-resource systems (e.g. city sizes bounded by total population, firm sizes bounded by total market size). Crucially, a power law with exponential cutoff has *all moments finite* (the exponential factor dominates at sufficiently large x), even though it can be visually indistinguishable from a pure power law over several orders of magnitude — a sobering reminder that *apparent* power-law behaviour over a finite observed range does not imply the formal (all-moments-divergent, Fréchet-domain) properties of a true power law extrapolated to infinity.

The "80/20" heuristic and the Lorenz curve. A power law with α close to 1 (the famous "80/20 rule" corresponds roughly to $\alpha \approx \log 5 / \log 4 \approx 1.16$ for the Pareto principle in its classic formulation) implies extreme concentration: a small fraction of observations account for a large fraction of the total. This is formalized by the *Lorenz curve* $L(p) = \frac{\int_0^{F^{-1}(p)} x f(x) dx}{\mathbb{E}[X]}$ and the *Gini coefficient* $G = 1 - 2\int_0^1 L(p)\, dp$. For a Pareto distribution with index $\alpha > 1$, $G = \frac{1}{2\alpha - 1}$ — as $\alpha \to 1^+$, $G \to 1$ (maximal inequality), while $\alpha \to \infty$ gives $G \to 0$ (equality). The Gini coefficient is thus a one-to-one (monotonic) reparametrization of the Pareto tail index for $\alpha > 1$, providing an interpretable summary of tail heaviness widely used in income/wealth distribution studies and, by extension, in any context where "concentration risk" (Chapter 4, Chapter 16) is of concern.

The Zipf law as a special case and the rank-frequency plot. Zipf's law (named for George Kingsley Zipf's 1949 observations on word frequencies) states that the r-th most frequent item has frequency proportional to r^{-1} — equivalently, the frequency decreases as a power law in rank. This is *precisely* a Pareto distribution with $\alpha = 1$ viewed from a rank-frequency perspective: if $N(\geq f) \sim f^{-1/\alpha}$ for a Pareto distribution with index α, then the r-th largest value satisfies $r \approx N(\geq x_r) \sim x_r^{-1/\alpha}$, giving $x_r \sim r^{-\alpha}$ — so rank-frequency Zipf exponent equals the Pareto tail index. The log-log rank-frequency plot (log frequency vs. log rank, or log value vs. log rank) is thus *equivalent* to the log-log CDF plot of Section 3.2, and all of Section 3.2's warnings about log-log OLS fitting apply equally to rank-frequency fits.

Example — Estimating α from rank-frequency data: the Zipf setting

For a corpus of N_{total} words (total word tokens) with V distinct types (vocabulary size), the frequency of the r-th most common word is $f_r \approx f_1 / r^{\beta}$ for some Zipf exponent $\beta \approx 1$ empirically. Taking $\hat{\alpha} = 1/\beta \approx 1$ — but the correct ML estimator requires care. If word counts n_w (number of times word w appears) are treated as i.i.d. Pareto with index α above a minimum frequency threshold $n_{\min}$ (the collection threshold, analogous to u in the GPD framework), the Aki-type estimator (Section E.5's formula from the seismology case study, which itself was a Pareto/exponential estimator in

magnitude units) gives $\hat{\alpha} = 1 + n_{\text{types}} \left[\sum_{w:n_w \geq n_{\min}} \log(n_w/n_{\min})\right]^{-1}$. For English text, $\hat{\alpha}$ is typically 0.9–1.1, remarkably close to 1 across many corpora — an empirical regularity whose deeper origins (cognitive/communicative optimization vs. statistical artifact of how languages are sampled and measured) remain debated. The closeness to $\alpha = 1$ (where the mean is barely finite, Section 3.1) is itself notable from a risk perspective: **word-frequency distributions, like income distributions and earthquake energies, sit close to the infinite-mean boundary**, a reminder that this mathematical boundary (between finite- and infinite-mean heavy tails) is not a mathematical abstraction but a real threshold crossed or approached by many empirical systems, with direct consequences for risk aggregation (Section 3.3 and 3.4).

Testing for power-law behaviour: the Vuong test in detail. Section 3.2 mentioned Vuong's test for comparing power-law to log-normal alternatives. The test statistic is

$$T_{\text{Vuong}} = \frac{\sum_{i=1}^{n} \ell_i}{\hat{\sigma}_\ell \sqrt{n}}, \qquad \ell_i = \log \frac{f_{\text{Pareto}}(x_i; \hat{\alpha}, \hat{x}_{\min})}{f_{\text{LN}}(x_i; \hat{\mu}, \hat{\sigma}_{\text{LN}})},$$

where f_{Pareto} and f_{LN} are, respectively, the fitted Pareto and log-normal densities (Section 3.2's "Step 4" methodology), and $\hat{\sigma}_\ell = \text{SD}(\ell_i)$ (sample standard deviation of the individual log-likelihood ratios). $T_{\text{Vuong}} \xrightarrow{d} \mathcal{N}(0,1)$ under the null of equal model fit, so $|T_{\text{Vuong}}| > 1.96$ at 5% significance indicates that one model is significantly better (positive T: Pareto preferred; negative T: log-normal preferred). A critical subtlety: **both the Pareto and log-normal fits should be done on the *same* restricted dataset** (observations above $\hat{x}_{\min}$, the same threshold for both), and with parameters fitted by the appropriate ML estimators for that restricted dataset (the log-normal MLE on the tail, restricted to $x \geq \hat{x}_{\min}$, is a truncated log-normal, not the ordinary log-normal MLE on all data). Fitting one model (Pareto) on the tail and the other (log-normal) on all the data produces a severely biased test, since the models are not being compared on a fair, equal footing.

3.3 Subexponential Distributions and the Principle of a Single Big Jump

Overview

This section addresses a question of central importance for risk aggregation: if $X_1, \ldots, X_n$ are i.i.d. heavy-tailed losses, what can be said about the tail of their *sum* $S_n = X_1 + \cdots + X_n$? The answer — for a broad class of distributions called *subexponential* — is the celebrated *principle of a single big jump*: the tail of S_n is asymptotically determined entirely by the probability that *one* of the X_i is itself large, not by the cumulative effect of many moderately large terms. This is the precise mathematical statement underlying the informal slogan "in heavy-tailed systems, the total is dominated by the largest single contribution."

Definition (Subexponential distribution)

A distribution F on $[0, \infty)$ is *subexponential* (written $F \in \mathcal{S}$) if, for X_1, X_2 i.i.d. $\sim F$,

$$\lim_{x \to \infty} \frac{P(X_1 + X_2 > x)}{P(X_1 > x)} = 2.$$

Interpretation. For *light*-tailed distributions (e.g. Exponential), $P(X_1 + X_2 > x) \sim \frac{x}{\theta} e^{-x/\theta} \cdot e^{-x/\theta}$-type behaviour is dominated by *both* variables being moderately large simultaneously

— the ratio $P(X_1+X_2>x)/P(X_1>x)\to\infty$ for the Exponential (the sum's tail decays *faster* relative to a single term — actually for Exponential, $P(X_1+X_2>x)\sim\frac{x}{\theta}e^{-x/\theta}$ while $P(X_1>x)=e^{-x/\theta}$, so the ratio $\to\infty$, NOT subexponential in this naive sense — illustrating that subexponentiality is a genuinely heavy-tail-specific property). For *subexponential* (typically heavy-tailed) distributions, by contrast, $P(X_1+X_2>x)\approx 2P(X_1>x)$ for large x: the event $\{X_1+X_2>x\}$ is, to leading order, just as likely to occur via "X_1 large, X_2 moderate" (probability $\approx P(X_1>x)$) as via "X_2 large, X_1 moderate" (probability $\approx P(X_2>x)$), with the event "both large" being asymptotically negligible by comparison.

Theorem (Principle of a single big jump)

If $F\in\mathcal{S}$, then for any fixed $n\geq 2$,

$$P(S_n>x)\sim n\,P(X_1>x)\qquad\text{as } x\to\infty,$$

where $S_n=X_1+\cdots+X_n$. Moreover, conditional on $\{S_n>x\}$ for large x, the dominant contribution comes from exactly one of the X_i being of order x, while the others remain $O(1)$:

$$P\left(\max_{i\leq n}X_i\leq x-y\,\middle|\,S_n>x\right)\to 0\qquad\text{for each fixed } y,\text{ as } x\to\infty.$$

Which distributions are subexponential? All regularly varying distributions ($\overline{F}\in RV_{-\alpha}$ for $\alpha>0$, Section 3.1) are subexponential — this covers the Pareto, Fréchet-domain, and Cauchy/Student-t families. The log-normal and Weibull distributions with shape parameter less than 1 are also subexponential, despite *not* being regularly varying — subexponentiality is strictly broader than regular variation, encompassing all distributions with tails decaying slower than any exponential ($\overline{F}(x)e^{\epsilon x}\to\infty$ for all $\epsilon>0$).

Example — Numerical illustration: sum of two Pareto random variables

Let X_1,X_2 i.i.d. Pareto($\alpha=2$, $x_{\min}=1$), so $\overline{F}(x)=x^{-2}$ for $x\geq 1$. We verify the single-big-jump principle by direct (approximate) computation. For large x,

$$P(X_1+X_2>x)=P(X_1>x)+P(X_2>x)+P(X_1\leq x,X_2\leq x,X_1+X_2>x)-[\text{overlap corrections}].$$

The first two terms sum to $2x^{-2}$. The third term — both variables moderate but summing to exceed x — can be shown to be $O(x^{-2}\log x)$ for $\alpha=2$ (a boundary case where there is a logarithmic correction; for $\alpha\neq$ integer values, this term is $o(x^{-\alpha})$, strictly lower order). So

$$P(X_1+X_2>x)\sim 2x^{-2}=2P(X_1>x),$$

confirming subexponentiality, with a logarithmic correction term that vanishes relative to x^{-2} but is non-negligible at finite (even quite large) x — a reminder that asymptotic statements like Karamata's theorem and the single-big-jump principle describe *limiting* behaviour, and finite-sample/finite-x corrections, while lower-order, can matter in practice.

Practical numbers. At $x=100$: $P(X_1>100)=10^{-4}$. The single-big-jump approximation gives $P(X_1+X_2>100)\approx 2\times 10^{-4}$. A more refined (numerically evaluated) calculation might give $\approx 2.3\times 10^{-4}$ — the 15% relative discrepancy is the logarithmic correction term, which would shrink (relatively) as x grows further, but is clearly not negligible at $x=100$.

Implications for risk aggregation and diversification. The single-big-jump principle has a dramatic and counter-intuitive consequence for portfolio risk: if individual losses $X_1, \dots, X_n$ are i.i.d. subexponential, the tail of the *total* loss S_n is, to leading order, n times the tail of a *single* loss — **diversification across n subexponential risks does not reduce the relative tail risk of the total compared to a single risk, in the extreme tail**. This stands in stark contrast to the light-tailed (e.g. Gaussian) case, where the CLT implies $\mathrm{Var}(S_n) = n\sigma^2$ but $S_n/n \to \mu$ (by LLN) with fluctuations of order $\sqrt{n}$ — diversification *does* reduce relative fluctuations for light tails.

> **The "diversification fallacy" in heavy-tailed risk.** A portfolio of n "independent" insurance policies, each with subexponential claim-size distribution, has *aggregate* extreme-loss probability that scales linearly in n (each policy contributes its own chance of "the big one"), not the $\sqrt{n}$ or faster reduction one might naively expect from a CLT/diversification intuition. This is the formal underpinning of why catastrophe insurance and reinsurance markets exist as they do: pooling many heavy-tailed risks does *not*, by itself, tame the tail — it merely *redistributes* the same aggregate tail risk across more contracts. True risk reduction in heavy-tailed settings requires either (a) genuine *negative dependence* or *bounding* mechanisms (caps, reinsurance layers, Section 4 risk measures with explicit tail treatment), or (b) reducing n (concentration) is *not* the answer either — it is the *tail index* α *itself*, a property of the underlying risk-generating process, that determines whether aggregation helps at all. This insight recurs throughout Chapters 4, 15, and 16 of this volume.

Closure properties. The class $\mathcal{S}$ is *not* closed under all the operations one might hope: subexponentiality of F does not imply subexponentiality of a mixture or a convolution with an arbitrary other distribution without additional conditions. However, $\mathcal{S}$ *is* closed under: (a) tail-equivalence (if $\overline{F}(x)/\overline{G}(x) \to c \in (0,\infty)$ and $G \in \mathcal{S}$, then $F \in \mathcal{S}$); (b) the class $RV_{-\alpha}$ for any $\alpha \geq 0$ is a subset of $\mathcal{S}$ (so all power-law and Fréchet-domain distributions are automatically subexponential); and (c) convolution within $RV_{-\alpha}$: if $F, G \in RV_{-\alpha}$ (same index), $F * G \in RV_{-\alpha}$ as well, with the single-big-jump asymptotic $\overline{F * G}(x) \sim \overline{F}(x) + \overline{G}(x)$.

The large deviations interpretation of the single-big-jump principle. The principle of a single big jump (Section 3.3) has an elegant interpretation in terms of large deviations theory (Chapter 17–18, Volume V of this series), which provides a complementary and in some respects deeper perspective.

For light-tailed F (Gumbel domain), the large deviations principle (Chapter 17) gives $P(S_n > nx) \sim e^{-nI(x)}$ for $x > \mathbb{E}[X]$, where $I(x) = \sup_t [tx - \log M(t)]$ is the Cramér rate function and $M(t) = \mathbb{E}[e^{tX}]$ is the moment generating function (finite for all $t > 0$ in the Gumbel domain). Crucially, the large-deviations event $\{S_n > nx\}$ is achieved, in this light-tailed regime, by *all n variables contributing approximately equally* — each $X_i \approx x$ rather than one being much larger than the others (the "democratic" mechanism of the CLT and law of large numbers). This is the exact *opposite* of the single-big-jump mechanism.

For heavy-tailed F (Fréchet domain, $\overline{F}(x) \sim Lx^{-\alpha}$), the moment generating function $M(t) = \mathbb{E}[e^{tX}] = \infty$ for all $t > 0$ — **the Cramér rate function does not exist**, large deviations theory in its classical form (Chapter 17) does not apply, and the probability $P(S_n > x)$ for x growing faster than linearly in n is *not* exponentially small in n (as it would be for light tails) but only polynomially small (since $P(S_n > x) \sim nP(X_1 > x) \sim nLx^{-\alpha}$). The single-big-jump principle is thus, from the large-deviations perspective, a statement about the *failure* of the classical Cramér-type large-deviations mechanism for heavy-tailed distributions — the tail is so heavy that the exponential tilting central to Cramér's theorem cannot be applied,

and instead the dominant mechanism for large-sum events is combinatorial (one term large) rather than collective (all terms slightly large).

Example — Exponential vs. polynomial tail decay: a quantitative comparison

For $X_1, \dots, X_{100}$ i.i.d., comparing $P(S_{100} > x)$ for two distributions with *the same mean* $\mathbb{E}[X] = 1$:
Light-tailed: $X \sim \text{Exp}(1)$ (mean 1, variance 1). Cramér bound: $P(S_{100} > x) \leq e^{-100I(x/100)}$ where $I(x) = x - 1 - \log x$ for $x > 1$ (Cramér rate for Exp(1)). At $x = 200$ (i.e. $x/100 = 2$): $I(2) = 2 - 1 - \log 2 = 1 - 0.693 = 0.307$, so $P(S_{100} > 200) \lesssim e^{-100 \times 0.307} = e^{-30.7} \approx 2.1 \times 10^{-14}$ — astronomically small.
Heavy-tailed: $X \sim \text{Pareto}(\alpha = 2, x_{\min} = 1)$ (mean $= 2/1 = 2$ — oops, different mean; instead use Pareto$(\alpha = 2, x_{\min} = 1/2)$, giving mean $= (2 \times 0.5)/(2-1) = 1$). Then $\overline{F}(x) = (x_{\min}/x)^\alpha = (0.5/x)^2 = 0.25/x^2$ for $x \geq 0.5$. Single-big-jump: $P(S_{100} > 200) \approx 100 \times P(X > 200) = 100 \times 0.25/200^2 = 25/40000 = 6.25 \times 10^{-4}$.
The ratio: $\frac{P_{\text{heavy}}(S_{100}>200)}{P_{\text{light}}(S_{100}>200)} \approx \frac{6.25 \times 10^{-4}}{2.1 \times 10^{-14}} \approx 3 \times 10^{10}$ — **the heavy-tailed sum exceeds the target** 200 **(twice the mean sum** $100 \times \mathbb{E}[X] = 100$**) about** 30 **billion times more frequently than the light-tailed sum with the same mean.** This stark numerical comparison captures, in a single ratio, the fundamental asymmetry between light-tailed and heavy-tailed risk: for a given distance above the mean (here, $2\times$ the mean), the heavy-tailed system exceeds it with a probability that is, for even modestly large n, orders of magnitude higher than any light-tailed system with the same mean — not by a constant factor, but by a factor growing exponentially in n (since the light-tailed probability decays exponentially while the heavy-tailed probability decays only polynomially).

3.4 Stable Distributions and Generalized Central Limit Theorems

Overview

The classical Central Limit Theorem requires finite variance and yields the Gaussian as the universal limit for normalized sums. When variance is infinite — as for Pareto distributions with $\alpha < 2$ (Section 3.1–3.3) — a *Generalized Central Limit Theorem* (GCLT) still holds, but the limiting distributions are the α*-stable* (Lévy stable) family, a richer class containing the Gaussian as the special case $\alpha = 2$. This section develops stable distributions, the GCLT, and their relationship to the heavy-tail and extreme-value theory of the preceding sections.

Definition (Stable distribution)

A random variable X has a *stable distribution* with index $\alpha \in (0, 2]$ if, for independent copies $X_1, X_2 \sim X$ and any $a, b > 0$, there exist $c > 0$ and $d \in \mathbb{R}$ such that

$$aX_1 + bX_2 \overset{d}{=} cX + d, \qquad \text{with } c^\alpha = a^\alpha + b^\alpha.$$

Equivalently, the characteristic function of a stable distribution has the form

$$\mathbb{E}[e^{itX}] = \exp\left\{i\mu t - |ct|^\alpha \left[1 - i\beta \,\text{sgn}(t)\, \omega(t, \alpha)\right]\right\},$$

where $\alpha \in (0, 2]$ is the *stability index*, $\beta \in [-1, 1]$ is the *skewness*, $c > 0$ is scale, μ is location, and $\omega(t, \alpha) = \tan(\pi\alpha/2)$ for $\alpha \neq 1$, $\omega(t, \alpha) = -\frac{2}{\pi} \log |t|$ for $\alpha = 1$.

Special cases. $\alpha = 2$ (with $\beta = 0$) gives the Gaussian (the only stable distribution with finite variance). $\alpha = 1, \beta = 0$ gives the Cauchy distribution. $\alpha = 1/2, \beta = 1$ gives the Lévy distribution (with an explicit closed-form density, used in first-passage-time problems, previewed in Volume II's discussion of degradation processes).

Theorem (Generalized Central Limit Theorem)

Let $X_1, X_2, \ldots$ be i.i.d. with $\overline{F}(x) \sim c_+ x^{-\alpha}$ and $F(-x) \sim c_- x^{-\alpha}$ as $x \to \infty$, for some $\alpha \in (0, 2)$ and $c_+, c_- \geq 0$ (not both zero) — i.e. symmetric or asymmetric Pareto-type tails with index $\alpha < 2$ (so variance is infinite). Then there exist $a_n > 0, b_n \in \mathbb{R}$ such that

$$\frac{S_n - b_n}{a_n} \xrightarrow{d} X_\alpha,$$

where X_α is α-stable with skewness $\beta = (c_+ - c_-)/(c_+ + c_-)$, and $a_n \sim n^{1/\alpha} L(n)$ for some slowly varying L.

For $\alpha = 2$ (finite variance), this reduces to the classical CLT with $a_n \sim \sqrt{n}$ and limit Gaussian. For $\alpha < 2$, $a_n \sim n^{1/\alpha}$, a *slower-growing* normalization than $\sqrt{n}$ is needed for the limit to be non-degenerate — equivalently, $S_n/n^{1/\alpha}$ converges (rather than $S_n/\sqrt{n}$), and since $1/\alpha > 1/2$ for $\alpha < 2$, S_n itself fluctuates on a *larger* scale than $\sqrt{n}$ would suggest: heavy tails produce larger-than-Gaussian fluctuations in sums.

Example — Numerical illustration: simulated sums of Pareto($\alpha = 1.5$) variables

For $\alpha = 1.5 < 2$, variance is infinite (recall from Section 3.1: moments exist only up to order α, so the second moment, $2 > 1.5$, diverges) but the mean is finite ($1 < 1.5$). The GCLT predicts $S_n/n^{1/1.5} = S_n/n^{2/3}$ converges to a 1.5-stable distribution (after centering).
Simulating $n = 1000$ sums each of $N = 10000$ i.i.d. Pareto($\alpha = 1.5, x_{\min} = 1$) variables (so each sum is itself a single draw, repeated 1000 times for the empirical distribution of S_N):

- $\mathbb{E}[S_N] \approx N \cdot \mathbb{E}[X] = 10000 \times 3 = 30000$ (since $\mathbb{E}[X] = \alpha x_{\min}/(\alpha - 1) = 1.5/0.5 = 3$ for Pareto(1.5,1)).
- Empirical std. dev. of S_N across the 1000 replications: if S_N were approximately Gaussian (CLT, finite variance assumption), we'd expect $\mathrm{SD}(S_N) \sim \sqrt{N}\sigma$ — but $\sigma = \infty$, so this is meaningless. Instead, $\mathrm{SD}(S_N)$ scales like $N^{1/\alpha} = N^{2/3} \approx 10000^{0.667} \approx 464$ *times* a stable-distribution scale parameter — in practice, the *empirical* standard deviation across replications will itself be highly variable and dominated by the occasional replication containing one extremely large X_i (the single-big-jump principle of Section 3.3 manifesting here too).
- A more robust diagnostic: the empirical *interquartile range* (IQR) of S_N across replications, scaled by $N^{1/\alpha}$, should stabilize as $N \to \infty$ — whereas the empirical variance of S_N across replications will *not* stabilize (it estimates an infinite population quantity) and will instead grow erratically with the number of replications, occasionally jumping when a replication contains an unusually extreme X_i.

This example illustrates a critically important practical point: **for $\alpha < 2$, sample variances and standard deviations of sums (or of the underlying data itself) are not merely "large"; they are statistically meaningless as estimates of a population quantity, because that population quantity is infinite.** Reporting a finite-sample standard

deviation for $\alpha < 2$ data, without flagging the tail index, is a common and serious error.

Stable distributions and extreme value theory: a comparison. It is important to distinguish the GCLT (governing *sums*, limit = stable distribution) from the FTG theorem (governing *maxima*, limit = GEV/Fréchet, Section 1.2). Both apply to heavy-tailed ($\alpha < \infty$) data and both involve the *same* tail index α, but describe different aggregation operations:

Aggregation	Normalization	Limit law
$M_n = \max(X_1, \dots, X_n)$	$a_n = n^{1/\alpha}$	Fréchet Φ_α
$S_n = X_1 + \cdots + X_n$ (finite var.)	$a_n = \sqrt{n}$	Gaussian
$S_n = X_1 + \cdots + X_n$ ($\alpha < 2$)	$a_n = n^{1/\alpha}$	α-stable

Strikingly, for $\alpha < 2$, *both* M_n and S_n require the *same* normalization $a_n = n^{1/\alpha}$ — a manifestation of the single-big-jump principle (Section 3.3): for heavy-tailed sums, the sum is dominated by its maximum term, so S_n and M_n have comparable order of magnitude, and indeed $S_n/M_n \to 1$ in an appropriate sense as the tail becomes heavier (smaller α).

Practical relevance: aggregation of operational risk losses. Operational risk losses in banking (fraud, system failures, legal settlements) are often modelled with $\alpha \in (0.5, 1.5)$ — i.e. *infinite mean* for $\alpha < 1$, or infinite variance for $\alpha < 2$. In such regimes, regulatory capital requirements based on aggregating a year's worth of losses via the GCLT must use α-stable (or related subexponential) aggregation, *not* a Gaussian/CLT-based approach — using the wrong aggregation law can underestimate required capital by orders of magnitude, particularly for α close to or below 1, where even the *mean* aggregate loss may be poorly estimated by a finite sample average (since the LLN itself can fail or converge pathologically slowly for $\alpha \leq 1$).

The Lévy–Khintchine representation and infinitely divisible distributions. The stable distributions of Section 3.4 are a special case of the more general class of *infinitely divisible* distributions, characterized by the Lévy–Khintchine theorem — a representation result of fundamental importance for the theory of random processes with independent increments (Lévy processes), which underpins many models in mathematical finance and risk theory.

Theorem (Lévy–Khintchine)

A random variable X is infinitely divisible (i.e. for every n, $X \overset{d}{=} X_1 + \cdots + X_n$ for i.i.d. X_i) if and only if its characteristic function has the form

$$\log \mathbb{E}[e^{itX}] = i\mu t - \frac{\sigma^2 t^2}{2} + \int_{\mathbb{R}\setminus\{0\}} \left(e^{itx} - 1 - \frac{itx}{1+x^2}\right) \nu(dx),$$

where $\mu \in \mathbb{R}$, $\sigma^2 \geq 0$, and ν (the *Lévy measure*) is a measure on $\mathbb{R} \setminus \{0\}$ satisfying $\int_{\mathbb{R}} \min(1, x^2)\nu(dx) < \infty$.

Connection to stable distributions. The α-stable distributions (Section 3.4) are precisely the infinitely divisible distributions whose Lévy measure has the form $\nu(dx) = \frac{c_+}{x^{1+\alpha}}dx$ for $x > 0$ and $\nu(dx) = \frac{c_-}{|x|^{1+\alpha}}dx$ for $x < 0$ (a power-law Lévy measure with index α). For $\alpha = 2$, $\nu \equiv 0$ (no jumps) and $\sigma^2 > 0$, recovering the Gaussian. For $\alpha < 2$, ν is non-zero (jumps present, with the jump-size distribution itself being a power law), and the Gaussian component σ^2 may be 0 (for purely discontinuous stable processes) or positive (for mixed

stable processes).

Example — The compound Poisson process as an infinitely divisible model for aggregate losses

The aggregate loss model $S = \sum_{i=1}^{N} X_i$ (with $N \sim \text{Poisson}(\lambda)$ and X_i i.i.d. severities, as in Appendix G) is infinitely divisible, with Lévy measure $\nu(dx) = \lambda f_X(x)dx$ (the Poisson rate times the severity density). This is the discrete-jump component of the Lévy–Khintchine representation; in the absence of a Gaussian component ($\sigma^2 = 0$), the compound Poisson process is a *pure-jump* Lévy process. When the severity distribution is heavy-tailed ($f_X(x) \sim Cx^{-\alpha-1}$, Section 3.1), the Lévy measure itself is regularly varying — $\nu((x,\infty)) \sim C\lambda x^{-\alpha}$ — and the resulting compound Poisson distribution for S inherits the single-big-jump property (Section 3.3): for large s, $P(S > s) \approx \lambda \int_s^\infty f_X(x)dx = \lambda P(X > s)$, consistent with Appendix G's Step 3 calculation.

Numerical connection. At the 99.9% level (Appendix G's $\text{VaR}_{0.999} \approx \$38.6\text{M}$), the compound-Poisson model predicts $P(S > \$38.6\text{M}) \approx \lambda P(X > \$38.6\text{M} - u)$, which at $\lambda = 0.92$ events/year above $u = \$0.5\text{M}$ (Appendix G's parameters) and $\hat{\xi} = 0.62, \hat{\beta} = \$0.68\text{M}$ gives $0.92 \times \overline{H}_{0.62,0.68}(38.1) \approx 0.92 \times 4.40 \times 10^{-4} \approx 4.05 \times 10^{-4}$ — i.e. the single-big-jump formula gives $P(S > \$38.6\text{M}) \approx 4.05 \times 10^{-4}$, slightly below the exact 0.001 $(= 10^{-3}$, the target $\text{VaR}_{0.999}$ level), confirming Appendix G's Step 3 finding (that the single-big-jump approximation at the 99.9% level gives $\approx 40\%$ of the exact probability, requiring "two-jump" corrections for full accuracy) from the Lévy-process perspective: the two-jump correction corresponds to the second-order term in the Lévy–Khintchine-based expansion of $P(S > s)$, involving $\int_0^s P(X_1 > s-x)dF_X(x)$ — the probability that two jumps together exceed s.

3.5 Multivariate Extremes and Tail Dependence

Overview

Real risks rarely occur in isolation: a market crash affects many asset prices simultaneously; an earthquake damages many structures in a region; a pandemic affects many countries. This final section of Chapter 3 extends extreme value theory to the multivariate setting, focusing on the crucial concept of *tail dependence* — the degree to which extreme events in one variable coincide with extreme events in another — and its profound implications for portfolio risk, which we will revisit extensively in Chapter 16's treatment of systemic risk.

Multivariate extreme value distributions. For a bivariate i.i.d. sample $(X_1, Y_1), \ldots, (X_n, Y_n)$, define componentwise maxima $M_n^X = \max_i X_i$, $M_n^Y = \max_i Y_i$ (note: these maxima need not occur for the same index i). Under appropriate normalization, the joint distribution of (M_n^X, M_n^Y) converges (if it converges at all) to a *multivariate extreme value (MEV) distribution*, whose margins are each (univariate) GEV (Section 1.3), but whose *dependence structure* is constrained to be *max-stable* — the multivariate analogue of the max-stability of Section 1.2.

Definition (Copula and tail dependence coefficient)

By Sklar's theorem, any joint CDF $H(x, y)$ with margins F, G can be written $H(x, y) = C(F(x), G(y))$ for a unique *copula* $C : [0,1]^2 \to [0,1]$ capturing the dependence struc-

ture independent of the margins. The *upper tail dependence coefficient* is

$$\lambda_U = \lim_{q \to 1^-} P\left(Y > G^{-1}(q) \mid X > F^{-1}(q)\right) = \lim_{q \to 1^-} \frac{1 - 2q + C(q,q)}{1-q}.$$

If $\lambda_U \in (0,1]$, X and Y are said to have (*upper*) *tail dependence*; if $\lambda_U = 0$, they are (*asymptotically*) *tail independent*.

The Gaussian copula and its (in)famous tail-independence property. A central, and for risk management a genuinely dangerous, fact is:

Theorem (Tail independence of the Gaussian copula)

For any bivariate Gaussian copula with correlation $\rho \in (-1,1)$ (i.e. $\rho \neq \pm 1$), $\lambda_U = 0$: **no matter how high the correlation ρ (short of perfect), the Gaussian copula exhibits zero tail dependence.**

Proof sketch. For bivariate standard normal (Z_1, Z_2) with correlation ρ, $P(Z_2 > z \mid Z_1 > z) \to 0$ as $z \to \infty$ for any $\rho < 1$, because the conditional distribution of $Z_2 \mid Z_1 = z$ is $\mathcal{N}(\rho z, 1-\rho^2)$ — as $z \to \infty$, $P(Z_2 > z \mid Z_1 = z) = P(\mathcal{N}(\rho z, 1-\rho^2) > z) = P(\mathcal{N}(0,1) > \frac{z(1-\rho)}{\sqrt{1-\rho^2}}) \to 0$ exponentially fast in z (since $\rho < 1$ means the exponent $\to \infty$). ■

Tail-dependent alternatives: the t-copula and extreme value copulas. The bivariate t-copula (derived from the multivariate Student-t distribution with ν degrees of freedom) has $\lambda_U = 2\overline{T}_{\nu+1}\left(\sqrt{(\nu+1)(1-\rho)/(1+\rho)}\right) > 0$ for any $\rho > -1$, where $\overline{T}_{\nu+1}$ is the Student-t survival function — the t-copula exhibits tail dependence that increases as ν decreases (heavier marginal tails) and as ρ increases.

Example — Numerical comparison: Gaussian vs t-copula tail dependence

For $\rho = 0.7$:

- Gaussian copula: $\lambda_U = 0$ exactly, for any sample size, in the limit.
- t-copula with $\nu = 4$: $\sqrt{(\nu+1)(1-\rho)/(1+\rho)} = \sqrt{5 \times 0.3/1.7} = \sqrt{0.882} \approx 0.940$. Then $\lambda_U = 2\overline{T}_5(0.940)$. Using $\overline{T}_5(0.940) \approx 0.196$ (Student-t with 5 df survival function at 0.94), $\lambda_U \approx 2 \times 0.196 \approx 0.392$.

So with the *same* linear correlation $\rho = 0.7$, the Gaussian model predicts that two assets *never* crash together in the extreme tail (asymptotically), while the t_4 model predicts they crash together with conditional probability $\approx 39\%$ in the extreme tail. **This single modelling choice — Gaussian vs. t copula, with identical correlation — changes the qualitative conclusion about joint extreme risk from "negligible" to "substantial."** This is widely regarded as one of the central modelling failures contributing to the underestimation of systemic risk prior to the 2007–2008 financial crisis (the so-called "Gaussian copula formula" used in CDO pricing).

Multivariate regular variation. A more general and powerful framework than copula-based tail dependence is *multivariate regular variation*: a random vector $\mathbf{X} = (X_1, \ldots, X_d)$ is multivariate regularly varying if there exists $\alpha > 0$ and a limiting measure ν on the unit sphere (or simplex) such that, for all Borel sets A bounded away from the origin,

$$\frac{P(\mathbf{X}/t \in A)}{P(|\mathbf{X}| > t)} \to \nu(A) \qquad \text{as } t \to \infty.$$

The *angular measure* ν encodes the full dependence structure of extremes: concentration of

ν near the "corners" of the simplex (e.g. near $(1,0)$ and $(0,1)$ in two dimensions) indicates tail *independence* (extremes tend to occur in only one component at a time), while concentration near the "diagonal" (e.g. near $(1/\sqrt{2}, 1/\sqrt{2})$) indicates tail *dependence* (extremes tend to occur simultaneously across components).

Implications for portfolio diversification under tail dependence. Section 3.3 established that diversification across i.i.d. heavy-tailed risks does not reduce relative tail risk (single-big-jump). The multivariate theory of this section adds a further layer: even when individual risks are *not* i.i.d. but *dependent*, the degree of *tail* dependence (not overall correlation) determines whether diversification helps in the extreme tail. Two assets can have low overall correlation ($\rho \approx 0.2$, say) yet substantial tail dependence (λ_U bounded away from 0) if their joint extremes share a common driver (e.g. both depend on a shared macroeconomic shock that is itself heavy-tailed) even though their day-to-day fluctuations are largely driven by idiosyncratic, less-correlated factors. **Overall correlation is therefore a poor and potentially deeply misleading proxy for tail risk co-movement** — a theme to which Chapter 4 (coherent risk measures) and Chapter 16 (systemic risk in financial networks) return repeatedly, and which motivates the explicit modelling of λ_U or the angular measure ν in any portfolio-level extreme value analysis.

Estimation of λ_U. A simple nonparametric estimator, based on the empirical copula, is

$$\hat{\lambda}_U(k) = \frac{1}{k} \sum_{i=1}^{n} \mathbf{1}\{R_i^X > n-k, R_i^Y > n-k\},$$

where R_i^X, R_i^Y are the ranks of X_i, Y_i in their respective samples, and k plays the role of a threshold (analogous to Section 2.3's threshold-selection problem, with the same bias-variance trade-offs and the same need for stability plots across choices of k). This estimator, and refinements based on multivariate regular variation, form the basis of empirical systemic risk indicators developed in Chapter 16.

PART 4

Risk Measures and Coherent Risk Theory

Formal axiomatics of risk measurement: Value-at-Risk, Conditional VaR, coherence axioms, and spectral risk measures.

The Pickands dependence function and its estimation. The upper-tail behaviour of a bivariate extreme value distribution (Section 3.5 and Appendix J.1) is fully characterized by a single function — the *Pickands dependence function* $A : [0,1] \to [1/2, 1]$ — which governs the joint extremes in a way analogous to how the scalar ξ governs univariate extremes.

Definition (Pickands dependence function)

For a bivariate extreme value distribution with standard Fréchet margins (i.e. margins transformed so $P(X_i \le x) = e^{-1/x}$ for $x > 0$, a natural normalization for the Fréchet domain), the joint survival function satisfies

$$P(X > x, Y > y) = \exp\left\{-(x+y)A\left(\frac{x}{x+y}\right)\right\}$$

for $x, y > 0$, where $A : [0,1] \to [1/2, 1]$ is the *Pickands dependence function*, satisfying $A(0) = A(1) = 1$ (boundary conditions), $A(t) \ge \max(t, 1-t)$ (dependence is intermediate between independence and perfect dependence), and A convex.

Special cases. $A(t) \equiv 1$ gives $P(X > x, Y > y) = e^{-(x+y)} = e^{-x}e^{-y}$ (independence). $A(t) = \max(t, 1-t)$ gives $P(X > x, Y > y) = e^{-\max(1/x, 1/y)}$ (perfect positive dependence, Fréchet upper bound). The Gumbel copula (Appendix J.1) has $A(t) = [t^\theta + (1-t)^\theta]^{1/\theta}$ for $\theta \geq 1$ — reducing to independence at $\theta = 1$ ($A \equiv 1$) and perfect dependence as $\theta \to \infty$.

The Pickands estimator

Given bivariate data $(X_1, Y_1), \ldots, (X_n, Y_n)$ with heavy-tailed margins, the Pickands nonparametric estimator of $A(t)$ is

$$\hat{A}_n(t) = \left[\frac{1}{n}\sum_{i=1}^{n} \min\left(\frac{1-t}{X_{(i)}}, \frac{t}{Y_{(i)}}\right)\right]^{-1}, \qquad t \in [0,1],$$

where $(X_{(i)}, Y_{(i)})$ are the data after marginal transformation to approximate standard Fréchet (via, e.g., the probability integral transform using empirical marginal CDFs). The estimator is consistent for $A(t)$ pointwise in t and, under regularity, asymptotically Gaussian with known (but complicated) variance function.

Example — Estimating $A(t)$ and testing independence

For the bivariate loss data of Appendix J.2–J.4 ($n = 200$, estimated Gumbel copula with $\hat{\theta} = 1.45$), the Pickands estimator $\hat{A}_n(t)$ at selected t values might give:

t	$\hat{A}_n(t)$	Gumbel-implied $A(t) = [t^{1.45} + (1-t)^{1.45}]^{1/1.45}$
0.25	0.839	0.832
0.50	0.772	0.763
0.75	0.838	0.832

The empirical values are close to the parametric Gumbel-implied values (differences of 0.007–0.009), consistent with the Gumbel copula fit of Appendix J.2–J.3.

Testing independence ($A(t) \equiv 1$): under independence, $n[1-\hat{A}_n(0.5)] \approx n[1-1] = 0$; any significantly positive $n[1-\hat{A}_n(0.5)]$ provides evidence against independence. Here $n[1-\hat{A}_n(0.5)] = 200 \times (1-0.772) = 200 \times 0.228 = 45.6$ — *strongly* rejecting independence ($\hat{A}_n(0.5) = 0.772$ is far below 1, and the asymptotic standard deviation of $\hat{A}_n(0.5)$ under independence is of order $n^{-1/2} \approx 0.07$, making the observed deviation from 1 about 3.3 standard deviations — highly significant). This nonparametric test for tail independence/dependence directly targets the quantity $\lambda_U = 2 - 2A(0.5)$ (the upper tail dependence coefficient, Section 3.5, recoverable from the Pickands function at $t = 0.5$ via $\lambda_U = 2 - 2A(1/2)$), providing a model-free test of tail independence that does not assume any specific copula family — an important complement to the parametric Gumbel-vs-Gaussian comparison of Appendix J.3, confirming tail dependence without requiring a parametric model to be correctly specified first.

4.1 Value-at-Risk (VaR) and Its Limitations

Overview

Value-at-Risk (VaR) is the most widely used — and most widely criticized — risk measure in finance and insurance. This section gives its formal definition in terms of quantiles (directly connecting to the return-level concepts of Section 2.5), derives its basic properties, and then systematically catalogues its mathematical limitations:

failure of subadditivity, insensitivity to tail shape beyond the quantile, and the resulting incentive problems it can create. These limitations motivate the coherent risk measure framework of Section 4.3 and the Expected Shortfall of Section 4.2.

Definition (Value-at-Risk)

For a loss random variable L (positive values denote losses) with CDF F_L, and confidence level $q \in (0,1)$ (typically $q = 0.95, 0.99, 0.999$), the *Value-at-Risk at level* q is

$$\mathrm{VaR}_q(L) = \inf\{x \in \mathbb{R} : F_L(x) \geq q\} = F_L^{-1}(q),$$

the q-quantile of the loss distribution.

Connection to return levels. If L represents the annual maximum loss, $\mathrm{VaR}_q(L)$ for $q = 1 - 1/T$ is exactly the T-year return level x_T of Section 2.5. The entire extreme value machinery of Chapters 1–3 is thus directly applicable to VaR estimation at high confidence levels (e.g. $q = 0.999$, corresponding to $T = 1000$).

Basic properties. VaR satisfies:

- *Monotonicity*: $L_1 \leq L_2$ (a.s.) $\Rightarrow \mathrm{VaR}_q(L_1) \leq \mathrm{VaR}_q(L_2)$.
- *Positive homogeneity*: $\mathrm{VaR}_q(\lambda L) = \lambda\,\mathrm{VaR}_q(L)$ for $\lambda > 0$.
- *Translation invariance*: $\mathrm{VaR}_q(L + c) = \mathrm{VaR}_q(L) + c$ for constant c.

These three properties — monotonicity, positive homogeneity, translation invariance — are shared by all the "coherent" risk measures discussed in Section 4.3. VaR's failure is in the *fourth* coherence axiom: subadditivity.

Theorem (Failure of subadditivity for VaR)

There exist random variables L_1, L_2 such that

$$\mathrm{VaR}_q(L_1 + L_2) > \mathrm{VaR}_q(L_1) + \mathrm{VaR}_q(L_2).$$

Example — Counterexample: VaR is not subadditive

Consider two independent loans, each defaulting independently with probability $p = 0.02$, with loss 0 if no default and loss 1 (in some unit) if default. Take $q = 0.95$.
For a single loan L_1: $F_{L_1}(0) = 0.98 \geq 0.95$, so $\mathrm{VaR}_{0.95}(L_1) = 0$. Similarly $\mathrm{VaR}_{0.95}(L_2) = 0$. Sum: $\mathrm{VaR}_{0.95}(L_1) + \mathrm{VaR}_{0.95}(L_2) = 0$.
For the combined portfolio $L_1 + L_2 \in \{0, 1, 2\}$: $P(L_1 + L_2 = 0) = (0.98)^2 = 0.9604$, $P(L_1 + L_2 \geq 1) = 1 - 0.9604 = 0.0396$. Since $0.9604 < 0.95 \leq 1$, the 0.95-quantile of $L_1 + L_2$ is 1 (the smallest value x with $F(x) \geq 0.95$ is $x = 1$, since $F(0) = 0.9604 < 0.95$ and $F(1) = 1 \geq 0.95$). So $\mathrm{VaR}_{0.95}(L_1 + L_2) = 1$.
We find $\mathrm{VaR}_{0.95}(L_1 + L_2) = 1 > 0 + 0 = \mathrm{VaR}_{0.95}(L_1) + \mathrm{VaR}_{0.95}(L_2)$. **VaR of the combined portfolio exceeds the sum of individual VaRs** — diversification, far from reducing risk as measured by VaR, appears to *increase* it. This is a direct consequence of VaR's discontinuous dependence on the tail shape: a small change in the probability of a rare event (0.02 here) can move the quantile discontinuously across an integer threshold.

The "cliff edge" problem: insensitivity beyond the quantile. A second, distinct, limitation:

$\text{VaR}_q(L)$ depends *only* on the q-quantile, and is completely insensitive to the *shape* of the distribution beyond that point. Two loss distributions can have identical $\text{VaR}_{0.99}$ but vastly different behaviour above the 99th percentile — one might have a sharp cutoff (bounded losses just above VaR) while the other has an extremely heavy tail (potential for catastrophic losses far beyond VaR), and VaR cannot distinguish them.

Example — Two distributions with identical VaR but different tail risk

Let $L_1 \sim$ a distribution with $P(L_1 > \text{VaR}_{0.99}) = 0.01$ and $L_1 \leq \text{VaR}_{0.99} + 1$ always (bounded excess). Let L_2 have the same $\text{VaR}_{0.99}$ but, conditional on exceeding it, follow a Pareto distribution with $\alpha = 1.2$ (infinite mean excess — recall Section 3.1, moments exist only up to order $\alpha < 1.2$, so even the *mean* of the excess is barely finite, and the variance is infinite).
$\text{VaR}_{0.99}(L_1) = \text{VaR}_{0.99}(L_2)$ by construction — a risk manager reporting only VaR would see *identical* risk for these two portfolios. Yet $\mathbb{E}[L_2 - \text{VaR}_{0.99}(L_2) \mid L_2 > \text{VaR}_{0.99}(L_2)]$ (the expected shortfall beyond VaR, Section 4.2) is bounded by 1 for L_1 but could be $10\times$, $100\times$, or more for L_2, depending on the Pareto scale parameter. **VaR is blind to exactly the information — the magnitude of losses in the tail beyond the threshold — that matters most for solvency and for understanding the consequences of a tail event actually occurring.**

Incentive effects: VaR and tail risk-seeking. Because VaR depends only on the probability of exceeding a threshold, not on the magnitude of the exceedance, an agent optimizing a portfolio subject to a VaR constraint has an incentive to take on *additional* tail risk *beyond* the VaR threshold, provided this does not increase the *probability* of exceedance — e.g. by writing far-out-of-the-money options or taking positions whose losses, conditional on a tail event, are unboundedly large but whose *probability* of any loss exceeding the VaR threshold is kept constant. This "VaR-gaming" incentive is one of the most serious criticisms of VaR-based regulation and is the primary motivation, alongside the subadditivity failure, for the shift toward Expected Shortfall in post-2008 banking regulation (Basel III/FRTB).

When is VaR subadditive? VaR *is* subadditive (in fact, coherent) for the special case of *elliptically distributed* losses (e.g. jointly Gaussian, or jointly Student-t with the same degrees of freedom for all components) — a fact sometimes used to defend VaR's continued practical use, since many financial models assume (at least approximately) elliptical or Gaussian return distributions. However, Section 3.5 already warned of the dangers of Gaussian (and even elliptical) dependence assumptions in the tail; precisely in the heavy-tailed, tail-dependent regimes where risk measurement matters most, the elliptical-distribution justification for VaR's subadditivity breaks down, and the counterexamples of this section become operative.

Historical context: the rise and institutionalization of VaR. Value-at-Risk's dominance in risk management is not purely a matter of mathematical convenience — it has a specific institutional and historical origin that illuminates both why VaR became universal and why its theoretical deficiencies were so long overlooked.

The decisive moment came in October 1994, when J.P. Morgan made public its internal "RiskMetrics" methodology, which had for several years been used internally to produce daily one-page risk summaries for senior management — the famous "4:15 reports" that presented a single dollar figure representing the potential daily loss at the 95% confidence level. The appeal of this reduction was enormous: instead of receiving risk reports describing portfolios in terms of Greeks (delta, gamma, vega, rho), scenario analyses, and sensitivity tables

requiring specialized expertise to interpret, non-specialist managers and board members could receive a single number: "we could lose at most $X today with 95% confidence." The communicative simplicity was the point, not the mathematical sophistication.

The institutionalization followed rapidly: the 1996 Amendment to the Basel Capital Accord formalized VaR (at the 99% level, 10-day holding period) as the regulatory standard for market risk capital, embedding it in binding international law. By the late 1990s, VaR was effectively mandatory for large financial institutions worldwide. The theoretical criticisms (Section 4.1's subadditivity failure was formally identified in the Artzner et al. 1997/1999 paper, Appendix H.2) arrived *after* institutionalization, and their translation into regulatory change took, as noted in Appendix H.2, roughly fifteen additional years.

A lesson in the sociology of risk measurement. VaR's institutionalization illustrates a general pattern: risk measurement methodologies, once embedded in regulatory frameworks and institutional practice, acquire substantial inertia independent of their theoretical merits, because the coordination costs of replacing them (retraining staff, repricing models, updating regulatory frameworks across dozens of jurisdictions simultaneously) are enormous. The mathematically superior alternative (CVaR/Expected Shortfall, Sections 4.2–4.3) was theoretically identified in the late 1990s, practically implementable by the early 2000s, and yet only formally adopted in major regulatory frameworks (FRTB) some two decades later. Understanding this inertia is part of understanding the real-world context in which the mathematics of this chapter operates — and part of the motivation for this volume's emphasis on *communicating* the mathematical properties (coherence, elicitability, uncertainty quantification) of risk measures clearly enough that their practical implications can be understood by the non-specialist audiences that ultimately determine which measures get institutionalized and which do not.

4.2 Conditional Value-at-Risk (CVaR) / Expected Shortfall

Overview

Conditional Value-at-Risk (CVaR), also called Expected Shortfall (ES) or Average Value-at-Risk (AVaR), addresses both limitations of VaR identified in Section 4.1: it is subadditive (hence coherent, Section 4.3), and it incorporates information about the *magnitude* of losses in the tail, not merely the probability of exceeding a threshold. CVaR has become the standard risk measure in post-2008 banking regulation (replacing VaR in the Basel Fundamental Review of the Trading Book) and is the natural risk measure to pair with the extreme value methodology of Chapters 1–3.

Definition (Conditional Value-at-Risk)

For a loss L with continuous distribution and confidence level $q \in (0,1)$,

$$\mathrm{CVaR}_q(L) = \mathbb{E}[L \mid L \geq \mathrm{VaR}_q(L)] = \frac{1}{1-q}\int_q^1 \mathrm{VaR}_u(L)\,du = \frac{1}{1-q}\mathbb{E}\left[L \cdot \mathbf{1}\{L \geq \mathrm{VaR}_q(L)\}\right].$$

For general (possibly discontinuous) distributions, the Rockafellar–Uryasev formula provides the correct general definition:

$$\mathrm{CVaR}_q(L) = \min_{c \in \mathbb{R}} \left\{ c + \frac{1}{1-q}\mathbb{E}[(L-c)^+] \right\},$$

and the minimizing $c^* = \mathrm{VaR}_q(L)$ recovers the quantile-based definition when F_L is continuous at $\mathrm{VaR}_q(L)$.

Connection to the mean excess function. For L with GPD-distributed excesses above a threshold $u = \mathrm{VaR}_q(L)$ (Section 2.1), $\mathrm{CVaR}_q(L) = \mathrm{VaR}_q(L) + e(\mathrm{VaR}_q(L))$ where $e(\cdot)$ is the mean excess function. Using the linear mean-excess formula of Section 2.1, $e(u) = \frac{\beta+\xi u}{1-\xi}$ (for $\xi < 1$):

$$\mathrm{CVaR}_q(L) = \mathrm{VaR}_q(L) + \frac{\beta + \xi\,\mathrm{VaR}_q(L)}{1-\xi} = \frac{\mathrm{VaR}_q(L) + \beta - \xi\,\mathrm{VaR}_q(L)\cdot 0}{1-\xi} \cdot (\ldots)$$

more directly:

$$\mathrm{CVaR}_q(L) = \frac{\mathrm{VaR}_q(L)}{1-\xi} + \frac{\beta - \xi u_0}{1-\xi} \qquad \text{(in terms of GPD parameters at a reference threshold } u_0\text{)}.$$

The key qualitative point: $\mathrm{CVaR}_q(L)$ **requires** $\xi < 1$ **to be finite** — if $\xi \geq 1$ (extremely heavy tails, infinite mean excess), CVaR itself is infinite, no matter how high q is chosen, a sharp and important contrast with VaR, which remains finite (it is just a quantile) for *any* ξ.

Example — Worked example: CVaR from a GPD tail model

Continuing the running example of Sections 2.4–2.5: threshold $u = 30$, $\hat{\xi} = 0.34$, $\hat{\beta} = 9.8$, $\hat{\lambda} = 4.5$ exceedances/year. Take $q = 0.99$ (annual basis), corresponding to $T = 100$ (since $1 - q = 0.01 = 1/100 = 1/T$).
From Section 2.5, $\mathrm{VaR}_{0.99} = x_{100} \approx 231.1$.
The mean excess at this level: $e(231.1) = \frac{\hat{\beta}+\hat{\xi}\times(231.1-30)}{1-\hat{\xi}}$ — using the threshold-stability property of Section 2.1, the scale parameter at threshold 231.1 is $\beta(231.1) = \hat{\beta} + \hat{\xi} \times (231.1 - 30) = 9.8 + 0.34 \times 201.1 \approx 9.8 + 68.4 = 78.2$. Then

$$e(231.1) = \frac{\beta(231.1)}{1-\hat{\xi}} = \frac{78.2}{0.66} \approx 118.5.$$

$$\mathrm{CVaR}_{0.99} = \mathrm{VaR}_{0.99} + e(\mathrm{VaR}_{0.99}) \approx 231.1 + 118.5 \approx 349.6.$$

CVaR exceeds VaR by more than 50% at this confidence level — the average loss, given that the 1-in-100-year threshold is breached, is roughly 349.6, compared to the threshold itself at 231.1. This large gap is itself informative: it reflects the heavy tail ($\hat{\xi} = 0.34$, well above 0), and would be much smaller for a light-tailed (Gumbel, $\xi \to 0$) loss distribution, where $e(u) \to \beta$ (constant) regardless of u.

CVaR as an integral of VaR: the spectral representation preview. The formula $\mathrm{CVaR}_q(L) = \frac{1}{1-q}\int_q^1 \mathrm{VaR}_u(L)\,du$ expresses CVaR as a *weighted average of VaRs at all levels above* q, with uniform weight $\frac{1}{1-q}$ on $u \in (q, 1)$ and zero weight below q. This is the simplest example of a *spectral risk measure* (Section 4.4), which generalizes CVaR by allowing non-uniform weighting functions — e.g. weighting extreme quantiles (u close to 1) more heavily than moderately extreme ones, to express greater risk aversion to truly catastrophic (vs. merely large) losses.

CVaR is coherent. Unlike VaR, CVaR satisfies all four coherence axioms (Section 4.3), including subadditivity: $\mathrm{CVaR}_q(L_1 + L_2) \leq \mathrm{CVaR}_q(L_1) + \mathrm{CVaR}_q(L_2)$ always. This follows from the Rockafellar–Uryasev variational formula: $\mathrm{CVaR}_q(L) = \min_c\{c + \frac{1}{1-q}\mathbb{E}[(L-c)^+]\}$ is a minimum of a family of functions, each of which is convex and subadditive in L (since

$(L-c)^+$ is convex in L and $\mathbb{E}[\cdot]$ is linear); a pointwise minimum of jointly convex functions in (L, c) is itself convex in L, and convexity plus positive homogeneity implies subadditivity.

Example — Revisiting the VaR counterexample with CVaR

Recall the two-loan example of Section 4.1: L_1, L_2 i.i.d., $P(L_i = 1) = 0.02$, $P(L_i = 0) = 0.98$, $q = 0.95$.
$\text{CVaR}_{0.95}(L_1) = \frac{1}{0.05}\mathbb{E}[(L_1 - c^*)^+$ adjustment]; more directly, using $\text{CVaR}_{0.95}(L_1) = \frac{1}{0.05}\mathbb{E}[L_1\mathbf{1}\{L_1 \geq \text{VaR}_{0.95}\}]$. Since $\text{VaR}_{0.95}(L_1) = 0$ (from Section 4.1) and $\{L_1 \geq 0\}$ is the entire sample space, $\text{CVaR}_{0.95}(L_1) = \frac{1}{0.05}\mathbb{E}[L_1] = \frac{0.02}{0.05} = 0.4$. Similarly $\text{CVaR}_{0.95}(L_2) = 0.4$. Sum: 0.8.
For L_1+L_2: $\text{VaR}_{0.95}(L_1+L_2) = 1$ (from Section 4.1). $\text{CVaR}_{0.95}(L_1+L_2) = \frac{1}{0.05}\mathbb{E}[(L_1+L_2)\mathbf{1}\{L_1 + L_2 \geq 1\}]$. We have $\mathbb{E}[(L_1 + L_2)\mathbf{1}\{L_1 + L_2 \geq 1\}] = 1 \times P(L_1 + L_2 = 1) + 2 \times P(L_1 + L_2 = 2)$. Compute $P(L_1 + L_2 = 1) = 2 \times 0.02 \times 0.98 = 0.0392$, $P(L_1 + L_2 = 2) = 0.02^2 = 0.0004$. So $\mathbb{E}[\cdot] = 1 \times 0.0392 + 2 \times 0.0004 = 0.04$. Then $\text{CVaR}_{0.95}(L_1 + L_2) = 0.04/0.05 = 0.8$.
$\text{CVaR}_{0.95}(L_1 + L_2) = 0.8 = \text{CVaR}_{0.95}(L_1) + \text{CVaR}_{0.95}(L_2)$ — subadditivity holds with *equality* in this case (reflecting the i.i.d., comonotonic-in-the-tail-region structure of this particular example), in sharp contrast to VaR's strict super-additivity $(1 > 0 + 0)$ in the same example.

Estimation challenges for CVaR. While CVaR has superior theoretical properties, it is statistically harder to estimate than VaR from a fixed sample: CVaR is an *average* over the tail beyond VaR, so its empirical estimator $\widehat{\text{CVaR}}_q = \frac{1}{\lceil n(1-q)\rceil}\sum_{i:L_i \geq \widehat{\text{VaR}}_q} L_i$ uses even fewer effective observations than the (already sparse) tail used for VaR, and inherits all the variance issues of Section 2.3's Hill-plot discussion — amplified, since CVaR is sensitive to the *largest* of these few observations (by the single-big-jump principle, Section 3.3, for heavy-tailed L). Backtesting CVaR (Section 4.5) is correspondingly more involved than backtesting VaR.

CVaR and the Neyman–Pearson lemma: a hidden optimality. There is a remarkable connection between CVaR and the classical Neyman–Pearson theory of hypothesis testing, which illuminates *why* CVaR arises naturally as the coherent measure that accompanies VaR.

Proposition (CVaR as optimal stress test)

Among all events A with $P(A) = 1 - q$ (i.e. sets of scenarios of "total probability" $1 - q$, the same probability mass as the VaR tail), the conditional expectation $\mathbb{E}[L \mid A]$ is *maximized* when $A = \{L \geq \text{VaR}_q(L)\}$ (the VaR exceedance set). Therefore

$$\text{CVaR}_q(L) = \max\{\mathbb{E}[L \mid A] : P(A) = 1 - q,\ A \in \mathcal{F}\}.$$

Proof. For any event A with $P(A) = 1 - q$, $\mathbb{E}[L \mid A] = \mathbb{E}[L\mathbf{1}_A]/(1 - q)$. We want to maximize $\mathbb{E}[L\mathbf{1}_A]$ over A with $P(A) = 1 - q$. By the Neyman–Pearson lemma (applied with L playing the role of the likelihood ratio), the set A that maximizes $\mathbb{E}[L\mathbf{1}_A]$ subject to $P(A) = c$ (a constant-probability constraint) is the set of L's c-largest values, i.e. $A^* = \{L \geq \text{VaR}_q(L)\}$ with $c = 1 - q$. ■

Interpretation. CVaR is not just *one* coherent alternative to VaR among many; it is the *worst-case conditional expectation over all events of the appropriate probability mass*. This gives CVaR a natural interpretation as a *stress test*: it is the answer to "what is the average loss, if we restrict attention to the worst fraction $1 - q$ of scenarios, chosen optimally to maximize the

average?" — and the optimal choice of that fraction turns out to be the VaR exceedance set, not some other set of scenarios. This Neyman–Pearson characterization also suggests a natural generalization: instead of restricting to $P(A) = 1 - q$ exactly, one could restrict to $P(A) \leq 1 - q$ (a "one-sided" constraint), which leads to a different (but related) optimization problem whose solution is connected to the *entropic risk measure* (Section 4.3), via the Lagrangian relaxation of the probability constraint.

Example — Stress scenario design using the Neyman–Pearson characterization

In regulatory stress testing (as opposed to VaR/CVaR calculation from a fitted model), practitioners often design "stress scenarios" — specific combinations of market moves (e.g. equity down 30%, credit spreads widen by 200bp, FX moves by 15%) intended to represent "severe but plausible" conditions. The Neyman–Pearson characterization suggests a principled approach to scenario design: **given a joint distribution for market risk factors and a portfolio loss function $L = L(\mathbf{r})$ (where $\mathbf{r}$ is the vector of risk-factor returns), the CVaR-maximizing event $A^* = \{L \geq \mathrm{VaR}_q(L)\}$ defines the set of risk-factor combinations that simultaneously (a) are individually consistent with the fitted risk-factor distribution (each scenario in A^* has positive density under the model) and (b) collectively account for the *worst* $1 - q$ fraction of outcomes**. The "centre of mass" of A^* — the conditional mean risk-factor vector $\mathbb{E}[\mathbf{r} \mid L(\mathbf{r}) \geq \mathrm{VaR}_q(L)]$ — can then serve as a model-consistent "representative stress scenario" that is specifically calibrated to maximize expected losses at the $1 - q$ tail level, rather than being chosen ad hoc based on historical analogy or expert judgment alone.

4.3 Coherent and Convex Risk Measures: Axiomatic Foundations

Overview

The coherent risk measure framework of Artzner, Delbaen, Eber, and Heath (1999) provides an axiomatic answer to the question "what properties *should* a sensible risk measure satisfy?" This section presents the four coherence axioms, shows CVaR satisfies them (completing the comparison with VaR begun in Sections 4.1–4.2), introduces the more general class of *convex* risk measures (relaxing positive homogeneity), and gives the representation theorem connecting coherent risk measures to "worst-case expectations" — directly linking this chapter to the robust optimization framework of Chapter 3 (Volume III of this series).

Definition (Coherent risk measure)

A risk measure $\rho : \mathcal{L} \to \mathbb{R}$ (mapping random losses to a real number, where $\mathcal{L}$ is some space of random variables on a probability space $(\Omega, \mathcal{F}, P)$) is *coherent* if it satisfies, for all $L, L_1, L_2 \in \mathcal{L}$ and $\lambda > 0$, $c \in \mathbb{R}$:

1. *Monotonicity*: $L_1 \leq L_2$ (a.s.) $\Rightarrow \rho(L_1) \leq \rho(L_2)$.
2. *Subadditivity*: $\rho(L_1 + L_2) \leq \rho(L_1) + \rho(L_2)$.
3. *Positive homogeneity*: $\rho(\lambda L) = \lambda \rho(L)$.
4. *Translation invariance*: $\rho(L + c) = \rho(L) + c$.

Interpretation of each axiom. Monotonicity is a basic sanity requirement: more loss in every scenario means more risk. Translation invariance says adding a sure loss c (e.g. a known cost) increases risk by exactly c — risk measures a *deviation/uncertainty* component on top

of the certain part. Positive homogeneity says doubling all exposures doubles the risk (no "diversification within a single position"). Subadditivity — the axiom VaR violates (Section 4.1) — says *combining* positions cannot increase risk beyond the sum of their individual risks; it is the formal expression of "diversification cannot hurt" (though, as Section 3.3 showed, for heavy-tailed risks it may not help much either — subadditivity permits, but does not guarantee, diversification benefit).

Definition (Convex risk measure)

A risk measure ρ is *convex* if it satisfies monotonicity, translation invariance, and *convexity*:

$$\rho(\lambda L_1 + (1-\lambda)L_2) \leq \lambda\rho(L_1) + (1-\lambda)\rho(L_2), \qquad \lambda \in [0,1].$$

Every coherent risk measure is convex (positive homogeneity + subadditivity $\Rightarrow$ convexity), but not conversely. Convex risk measures allow for *liquidity risk* effects: doubling a position might increase risk by *more* than double (e.g. due to market impact when liquidating a large position), violating positive homogeneity while preserving convexity.

Theorem (Robust representation of coherent risk measures)

ρ is a coherent risk measure (on a suitable space, with appropriate continuity) if and only if there exists a set $\mathcal{Q}$ of probability measures (absolutely continuous w.r.t. P) such that

$$\rho(L) = \sup_{Q \in \mathcal{Q}} \mathbb{E}_Q[L].$$

For convex (not necessarily coherent) risk measures, the representation generalizes to

$$\rho(L) = \sup_{Q \in \mathcal{Q}} \left(\mathbb{E}_Q[L] - \alpha(Q) \right)$$

for some "penalty function" $\alpha : \mathcal{Q} \to [0, \infty]$.

Interpretation: risk as worst-case expected loss. The representation theorem says: a coherent risk measure is the *worst-case expected loss* over a family $\mathcal{Q}$ of "stress" probability measures (alternative models for the data-generating process), and a convex risk measure is a worst-case *penalized* expected loss, where the penalty $\alpha(Q)$ reflects how "implausible" the alternative measure Q is (smaller penalty = more plausible alternative, contributing more to the worst case). This is *exactly* the structure of the robust and distributionally robust optimization problems that will be developed systematically in Volume III (Chapters 9–10): **coherent risk measurement and robust decision-making are, formally, the same mathematical object**, viewed from different angles.

Example — The representation set $\mathcal{Q}$ for CVaR

For CVaR_q, the representing set is

$$\mathcal{Q}_{\text{CVaR}_q} = \left\{ Q : \frac{dQ}{dP} \leq \frac{1}{1-q} \text{ a.s.} \right\},$$

i.e. all probability measures whose Radon–Nikodym derivative w.r.t. the physical measure P is bounded by $\frac{1}{1-q}$. Intuitively, $\mathcal{Q}_{\text{CVaR}_q}$ consists of measures that can "reweight" probability mass toward the worst $(1-q)$-fraction of outcomes, but no further — $\text{CVaR}_q(L) = \sup_{Q \in \mathcal{Q}_{\text{CVaR}_q}} \mathbb{E}_Q[L]$ is achieved by the Q^* that places *all* available extra

weight (up to the bound $\frac{1}{1-q}$) on the worst $(1-q)$ fraction of scenarios for L — exactly recovering $\mathrm{CVaR}_q(L) = \mathbb{E}[L \mid L \geq \mathrm{VaR}_q(L)]$.
For VaR_q: **no such representation exists** — there is no set $\mathcal{Q}$ of probability measures such that $\sup_{Q\in\mathcal{Q}} \mathbb{E}_Q[L] = \mathrm{VaR}_q(L)$ for all L, because $\sup_Q \mathbb{E}_Q[L]$ is necessarily subadditive in L (as a supremum of linear — hence subadditive — functionals), while VaR_q is not subadditive (Section 4.1). **This is the deepest reason VaR is not coherent**: it cannot be expressed as a worst-case expectation over any family of alternative scenarios, and therefore does not correspond to any coherent notion of "planning for the worst case within a plausible set of models."

Entropic and other convex (non-coherent) risk measures. The *entropic risk measure*, $\rho_\gamma(L) = \frac{1}{\gamma}\log \mathbb{E}[e^{\gamma L}]$ for risk-aversion parameter $\gamma > 0$, is convex but not positively homogeneous ($\rho_\gamma(\lambda L) \neq \lambda\rho_\gamma(L)$ in general), hence not coherent — yet it has a natural representation with penalty function $\alpha(Q) = \frac{1}{\gamma}D_{\mathrm{KL}}(Q\|P)$ (the Kullback–Leibler divergence), connecting it to the ϕ-divergence ambiguity sets of distributionally robust optimization (Section 10.1 of Volume III) and to large deviations theory (Chapter 17–18 of this volume, Part V) — the entropic risk measure is, in fact, exactly the large-deviations rate-function-based "tilted" expectation that appears throughout Cramér's theorem and its generalizations.

Law-invariance and the Kusuoka representation. A risk measure is *law-invariant* if $\rho(L)$ depends only on the distribution of L (not on the underlying probability space) — a natural requirement for any risk measure used to compare different random variables. Kusuoka's theorem (2001) shows that every law-invariant coherent risk measure on a suitable space can be represented as

$$\rho(L) = \sup_{\mu\in\mathcal{M}} \int_0^1 \mathrm{CVaR}_q(L)\,\mu(dq)$$

for some set $\mathcal{M}$ of probability measures on $[0,1]$ — **every law-invariant coherent risk measure is (a worst-case mixture of) integrals of CVaR at various confidence levels**. This remarkable result shows that CVaR is not merely *one* coherent alternative to VaR, but in a precise sense the *fundamental building block* from which all law-invariant coherent risk measures are constructed — directly motivating the spectral risk measures of Section 4.4, which correspond to the special case $|\mathcal{M}| = 1$ (a single, non-worst-case mixing measure μ).

Risk measures and no-arbitrage pricing: the dual connection. The robust representation of coherent risk measures (Section 4.3's theorem, $\rho(L) = \sup_{Q\in\mathcal{Q}} \mathbb{E}_Q[L]$) is structurally identical to the *no-arbitrage pricing* formula from mathematical finance, where the price of a derivative payoff H is

$$\Pi(H) = \sup_{Q\in\mathcal{Q}_{\mathrm{EMM}}} \mathbb{E}_Q[H]$$

(the supremum over *equivalent martingale measures* $\mathcal{Q}_{\mathrm{EMM}}$, the set of risk-neutral probability measures consistent with no-arbitrage, in an incomplete market where the pricing measure is not unique). **A coherent risk measure on a portfolio of loss claims is mathematically equivalent to an upper pricing rule in an incomplete market**: both are suprema of linear (expectation) functionals over a convex set of probability measures, and both preserve the key economic properties of monotonicity (larger payoff ⇒ higher price/higher risk), translation invariance (adding a bond reduces risk/price by exactly the bond's face value), and

subadditivity (portfolio diversification cannot increase risk/cost above the sum of individual risks/costs).

The superreplication dual and its risk-measure interpretation

In an incomplete market, the *superreplication price* of a contingent claim H is the minimum initial capital p^* needed to replicate H in every state ("superhedge"), given by (by duality theory)

$$p^* = \sup_{Q \in \mathcal{Q}_{\text{EMM}}} \mathbb{E}_Q[H].$$

This is precisely the coherent risk measure representation with $\mathcal{Q} = \mathcal{Q}_{\text{EMM}}$. **The capital required (by a coherent risk measure) to make a position safe is, mathematically, the same as the cost to superhedge that position in the incomplete-market sense** — the set of "plausible alternative models" $\mathcal{Q}$ in the risk measure plays the role of the set of equivalent martingale measures in pricing.

This connection is more than formal: it suggests that coherent risk measures and no-arbitrage pricing are two faces of the same duality, and that techniques developed in one context transfer to the other. For example:

- The *dual problem* of minimizing capital subject to a risk constraint (a portfolio optimization problem in the risk-measure context) corresponds to the *primal problem* of finding the superreplicating portfolio (a hedging problem in the pricing context) — the two are related by LP/convex duality.
- *Model ambiguity* (choosing $\mathcal{Q}$ to represent uncertainty about the loss model, as in the distributionally robust optimization of Volume III's Chapter 10) corresponds to *market incompleteness* (the set of martingale measures is large when markets are incomplete, reflecting genuine ambiguity about the "correct" pricing measure) — the mathematical treatment is identical, even though the economic interpretation differs.

Example — CVaR as an incomplete-market price

If $\mathcal{Q}_{\text{EMM}}$ is the set of probability measures with Radon–Nikodym derivative $dQ/dP \in [0, 1/(1-q)]$ (the same set characterizing CVaR_q, Section 4.3), then $\text{CVaR}_q(H) = \sup_{Q \in \mathcal{Q}_{\text{EMM}}} \mathbb{E}_Q[H]$ is simultaneously (a) the CVaR of the loss H, and (b) the superreplication price of H in a market where the only martingale measures are those with bounded density ratio $\leq 1/(1-q)$ relative to the historical measure P. **The CVaR confidence level q thus directly controls the degree of market incompleteness**: higher q (tighter risk measure, larger capital requirement) corresponds to a smaller set of martingale measures (more constrained market, harder to hedge), with the Sharpe ratio / market price of risk of the "worst-case" measure Q^* (the measure achieving the supremum) directly related to q through the density bound $1/(1-q)$.

4.4 Spectral and Distortion Risk Measures

Overview

Spectral risk measures generalize CVaR by allowing an arbitrary weighting ("spectrum") of VaRs across confidence levels, subject to a monotonicity condition that guarantees coherence. This section develops the spectral framework, its equivalent formulation as a *distortion* of the underlying probability measure, and discusses how the choice of spectrum encodes a decision-maker's risk preferences — connecting risk measurement to the decision theory under uncertainty developed in Chapter 11 (Volume III).

Definition (Spectral risk measure)

Given a non-negative, non-decreasing, integrable weight function $\phi : [0,1] \to [0,\infty)$ with $\int_0^1 \phi(u)\,du = 1$ (the *risk spectrum*), the associated spectral risk measure is

$$\rho_\phi(L) = \int_0^1 \phi(u)\,\mathrm{VaR}_u(L)\,du.$$

Recovering known measures. $\phi(u) = \frac{1}{1-q}\mathbf{1}\{u \geq q\}$ recovers CVaR_q (uniform weight on $[q,1]$, zero below). $\phi(u) \equiv 1$ recovers $\rho_\phi(L) = \mathbb{E}[L]$ (the risk-neutral expectation — the "flattest" possible spectrum). Any ϕ that is non-decreasing and integrates to 1 assigns *more weight to higher quantiles* (worse outcomes), in a graduated rather than all-or-nothing (CVaR-style) manner.

Theorem (Coherence of spectral risk measures)

ρ_ϕ is coherent for any non-negative, non-decreasing ϕ with $\int_0^1 \phi = 1$.

Proof sketch. Translation invariance and positive homogeneity follow directly from the corresponding properties of VaR_u and linearity of the integral. Monotonicity follows from monotonicity of VaR_u in L (for each u) and $\phi \geq 0$. Subadditivity is the only non-trivial part, and follows from the fact that ρ_ϕ can itself be written as a mixture (via the Kusuoka representation of Section 4.3) of CVaR_q's, each of which is subadditive: specifically, ϕ non-decreasing implies $\phi(u) = \int_0^1 \frac{1}{1-q}\mathbf{1}\{u \geq q\}\,d\nu(q)$ for some positive measure ν (essentially, ϕ is a mixture of the CVaR-spectra $\frac{1}{1-q}\mathbf{1}\{\cdot \geq q\}$, weighted by $d\nu(q) = -(1-q)\,d\phi(q)$ via integration by parts), so $\rho_\phi(L) = \int_0^1 \mathrm{CVaR}_q(L)\,d\nu(q)$, a non-negative mixture of subadditive functionals, hence subadditive. ■

The distortion function formulation. An equivalent and often more intuitive formulation defines a *distortion function* $g : [0,1] \to [0,1]$, non-decreasing, $g(0) = 0, g(1) = 1$, concave (the concavity corresponds to $\phi = g'$ being non-increasing in the *survival* parametrization — conventions vary in the literature; we adopt the convention where concave g applied to the survival function gives more weight to tail events), and define

$$\rho_g(L) = \int_0^\infty g(\overline{F}_L(x))\,dx - \int_{-\infty}^0 [1 - g(\overline{F}_L(x))]\,dx,$$

where $\overline{F}_L = 1 - F_L$. For $L \geq 0$, this simplifies to $\rho_g(L) = \int_0^\infty g(\overline{F}_L(x))\,dx$ — the expectation of L computed under the "distorted" survival function $g \circ \overline{F}_L$ rather than $\overline{F}_L$ itself. Since

$\mathbb{E}[L] = \int_0^\infty \overline{F}_L(x)\,dx$ for $L \geq 0$, ρ_g is literally $\mathbb{E}[L]$ computed after *distorting* the probabilities — inflating the probability mass assigned to large x (since g concave with $g(0) = 0, g(1) = 1$ implies $g(p) \geq p$ for $p \in [0,1]$, so $g(\overline{F}(x)) \geq \overline{F}(x)$, i.e. the distorted survival function is everywhere at least as large as the true one, assigning *more* probability to $\{L > x\}$ for every x).

Example — The proportional hazards (power) distortion family

A widely used family is $g_\gamma(p) = p^{1/\gamma}$ for $\gamma \geq 1$ (the "proportional hazards transform"). $\gamma = 1$ gives $g(p) = p$ (no distortion, $\rho_g = \mathbb{E}[L]$). For $\gamma > 1$, $g_\gamma(p) = p^{1/\gamma} > p$ for $p \in (0,1)$ (concave, since $1/\gamma < 1$), increasingly so as γ grows — larger γ means more risk-averse.

Numerical illustration. Suppose L ∼GPD-tailed with $\overline{F}_L(x) = (1 + \xi x/\beta)^{-1/\xi}$ for $x \geq 0$ (a purely Pareto-type tail for simplicity, ignoring the bulk). With $\xi = 0.3, \beta = 10$, and $\gamma = 2$:

$$\rho_{g_2}(L) = \int_0^\infty \left[(1 + 0.3x/10)^{-1/0.3}\right]^{1/2} dx = \int_0^\infty (1+0.03x)^{-1/0.6}\,dx = \int_0^\infty (1+0.03x)^{-1.667}\,dx.$$

This integral evaluates to $\frac{1}{0.03\times(1.667-1)} = \frac{1}{0.03\times 0.667} \approx 50.0$.
Compare to $\mathbb{E}[L] = \int_0^\infty (1+0.03x)^{-1/0.3} dx = \int_0^\infty (1+0.03x)^{-3.333} dx = \frac{1}{0.03\times(3.333-1)} = \frac{1}{0.03\times 2.333} \approx 14.3$.
So $\rho_{g_2}(L) \approx 50.0$ versus $\mathbb{E}[L] \approx 14.3$ — the distorted (risk-adjusted) expectation is more than $3.5\times$ the actual expected loss, reflecting $\gamma = 2$'s substantial loading for tail risk. As γ increases further, $\rho_{g_\gamma}(L)$ increases without bound and eventually becomes infinite once $1/(\gamma\xi) \leq 1$, i.e. $\gamma \geq 1/\xi = 1/0.3 \approx 3.33$ — beyond this point, the distorted tail $(1 + \xi x/\beta)^{-1/(\gamma\xi)}$ decays too slowly (exponent ≤ 1) for the integral to converge, mirroring the $\xi \geq 1$ infinite-CVaR phenomenon of Section 4.2 but now triggered by the *distortion parameter* γ rather than ξ alone.

The Wang transform and dual utility theory. Spectral/distortion risk measures have a decision-theoretic interpretation via *dual utility theory* (Yaari, 1987): rather than transforming *outcomes* via a utility function $u(\cdot)$ (as in expected utility theory) and taking $\mathbb{E}[u(L)]$, dual utility transforms *probabilities* via a distortion g and computes the expectation under the distorted measure. This is formally equivalent to *rank-dependent expected utility* with a particular structure, and provides an axiomatic decision-theoretic foundation for spectral risk measures — an agent whose preferences satisfy a certain set of axioms (weaker than the von Neumann–Morgenstern axioms, in particular relaxing the independence axiom) necessarily evaluates risky prospects via some distortion g, with ρ_g as the corresponding risk measure for losses.

Choosing a spectrum in practice. The flexibility of spectral risk measures — any non-decreasing ϕ integrating to 1 gives a coherent measure — is both a strength and a weakness: it provides a rich family for expressing graduated risk aversion (e.g. a regulator might specify ϕ to penalize losses beyond the 99.5% quantile increasingly steeply, rather than the all-or-nothing cutoff of CVaR), but the choice of ϕ is itself a policy/preference decision not dictated by the data. In practice, ϕ is often chosen from a low-dimensional parametric family (such as the power distortion g_γ above, or the Wang transform $g(p) = \Phi(\Phi^{-1}(p) + \lambda)$ for a single "market price of risk" parameter λ), calibrated to match observed insurance or option prices (which themselves

embed a market-implied risk distortion) — an approach that connects risk measurement directly to asset pricing theory, a connection explored further in the context of systemic risk pricing in Chapter 16.

The Esscher transform and premium calculation in insurance. The spectral risk measure framework (Section 4.4) has a long history in actuarial science, predating its modern mathematical formulation, in the form of the *Esscher transform* — a classical method for distorting probabilities in insurance premium calculation.

Definition (Esscher transform and Esscher premium)

For a loss L with moment generating function $M_L(h) = \mathbb{E}[e^{hL}] < \infty$ for $h > 0$, the *Esscher transform* with parameter h is the probability measure Q_h with $dQ_h/dP = e^{hL}/M_L(h)$. The *Esscher premium* is $\Pi_h(L) = \mathbb{E}_{Q_h}[L] = M_L'(h)/M_L(h) = \frac{d}{dh}\log M_L(h)$.

Connection to the entropic risk measure (Section 4.3). The Esscher premium can be written $\Pi_h(L) = \rho_{\text{ent},h}(L)$ where $\rho_{\text{ent},h}$ is (a variant of) the entropic risk measure with parameter h, whose convex risk measure representation (Section 4.3's general convex representation) uses penalty $\alpha(Q) = (1/h)D_{\text{KL}}(Q\|P)$ — i.e. the Esscher transform is the *exponential tilting* that solves the entropic risk minimization problem. However, unlike the coherent risk measures of Section 4.3, the Esscher premium requires $M_L(h) < \infty$ — ruling out application to Fréchet-domain (heavy-tailed) losses where $M_L(h) = \infty$ for all $h > 0$. **This is precisely the limitation of the Esscher/entropic approach for the heavy-tailed losses that are the focus of this volume**: any premium calculation requiring a finite MGF (and hence any large-deviation-rate-function-based calculation of the type developed in Volume V) is, formally speaking, inapplicable to genuinely heavy-tailed (Fréchet-domain) losses. The spectral risk measures of Section 4.4, which do not require a finite MGF (they are defined purely via the quantile function), provide the appropriate framework for heavy-tailed insurance and financial losses, which is one reason they have increasingly displaced the Esscher transform in modern actuarial practice for catastrophe (re)insurance pricing, where Fréchet-domain severity distributions (as in Appendix E.2's case study, with $\hat{\xi} \approx 0.45$) are the norm rather than the exception.

Example — Comparing Esscher and distortion premiums for a Pareto loss

Let $L \sim \text{Pareto}(\alpha = 3, x_{\min} = 1)$, so $\mathbb{E}[L] = \alpha x_{\min}/(\alpha - 1) = 1.5$, $\text{Var}(L) = \alpha x_{\min}^2/[(\alpha - 1)^2(\alpha - 2)] = 3/(4 \times 1) = 0.75$, $M_L(h) = \infty$ for all $h > 0$ (since $P(L > x) \sim x^{-3}$ is regularly varying, not exponentially decaying). **The Esscher premium is undefined for all** $h > 0$ — the Pareto distribution, despite having finite mean and variance, has no finite MGF, hence no Esscher transform.

The Wang-distortion premium (Section 4.4's proportional hazards/power distortion family) at $\gamma = 2$: $\rho_{g_2}(L) = \int_0^\infty g_2(\overline{F}_L(x))dx = \int_1^\infty (x^{-3})^{1/2}dx = \int_1^\infty x^{-3/2}dx = [x^{-1/2}/(-1/2)]_1^\infty = 0 - (-2) = 2$. So the Wang premium is $\rho_{g_2}(L) = 2$ versus the pure expected value $\mathbb{E}[L] = 1.5$ — a 33% safety loading, entirely driven by the tail-risk component captured by the distortion g_2, without requiring any finite-MGF assumption. The distortion framework thus provides what the Esscher/entropic framework cannot: a **well-defined, coherent, tail-sensitive premium for heavy-tailed losses**, a fact that is rapidly becoming standard actuarial knowledge but is still not universally reflected in industry premium calculation practice, where various MGF-based approx-

imations (often justified by appeal to log-normal or similar "thin-enough" tail models) continue to be used even for genuinely heavy-tailed perils.

4.5 Backtesting and Elicitability of Risk Measures

Overview

This closing section of Chapter 4 addresses a question with significant regulatory and practical weight: how does one *validate*, from realized data, whether a risk measure forecast was accurate — i.e. *backtest* it? We show that VaR, despite its theoretical deficiencies (Section 4.1), is straightforwardly backtestable, while CVaR — despite its theoretical superiority (Section 4.2) — poses a subtle and important statistical challenge: CVaR alone is *not elicitable*. We resolve this via joint elicitability of the pair (VaR, CVaR), and discuss the practical backtesting procedures used in banking regulation.

Backtesting VaR: the binomial/traffic-light approach. If $\mathrm{VaR}_q(L_t)$ is forecast for each day $t = 1, \dots, N$, and L_t is the realized loss, define the *violation indicator* $I_t = \mathbf{1}\{L_t > \mathrm{VaR}_q(L_t)\}$. If the forecasts are correctly calibrated, $I_t \sim \text{Bernoulli}(1-q)$ i.i.d., so $\sum_t I_t \sim \text{Binomial}(N, 1-q)$ under the null hypothesis of correct calibration.

Example — Numerical backtesting example

For $q = 0.99$ and $N = 250$ trading days (one year), the expected number of violations is $N(1-q) = 2.5$. The Basel "traffic light" framework classifies outcomes:

Zone	Violations	Approx. probability under H_0 (correct calibration)
Green	0–4	$\approx 89\%$
Yellow	5–9	$\approx 10.5\%$
Red	≥ 10	$\approx 0.5\%$

(Probabilities computed from $\text{Binomial}(250, 0.01)$: $P(X \leq 4) \approx 0.8907$, $P(5 \leq X \leq 9) \approx 0.1050$, $P(X \geq 10) \approx 0.0043$.) A bank in the "red zone" (10+ violations in 250 days, more than $4\times$ the expected number) faces escalating capital multipliers under Basel rules — this is a direct, simple, and *elicitable* (see below) backtest, one of the reasons VaR remained the regulatory standard for so long despite its theoretical issues.

Conditional coverage and independence tests. The binomial test above checks only the *unconditional* violation rate; it does not detect *clustering* of violations (e.g. all 10 violations occurring in the same volatile week, versus spread evenly through the year) which would indicate the VaR model fails to adapt to changing volatility. The Christoffersen (1998) test extends the binomial test to jointly assess (a) correct unconditional coverage and (b) independence of consecutive violation indicators I_t, I_{t+1}, via a likelihood ratio test comparing the observed transition frequencies $\{I_t = i \to I_{t+1} = j\}$ to those expected under independence.

Definition (Elicitability)

A statistical functional Θ (mapping a distribution F to a real number $\Theta(F)$) is *elicitable* if there exists a *scoring function* $S(x, y) \geq 0$ such that

$$\Theta(F) = \arg\min_{x \in \mathbb{R}} \mathbb{E}_{Y \sim F}[S(x, Y)].$$

VaR is elicitable. $\text{VaR}_q(F)$ is elicitable with the *pinball loss* (quantile loss) scoring function

$$S_q(x, y) = (q - \mathbf{1}\{y < x\})(y - x) = \begin{cases} q(y - x), & y \geq x, \\ (1 - q)(x - y), & y < x. \end{cases}$$

This is the loss function underlying quantile regression; $\arg\min_x \mathbb{E}[S_q(x, Y)] = F^{-1}(q)$, the q-quantile, exactly VaR_q. Elicitability via this scoring function is what makes the simple violation-counting backtest of the example above valid: the expected score is minimized at the true VaR, so systematic deviations of the forecast from the truth are detectable via the empirical violation rate.

Theorem (Gneiting, 2011: CVaR is not elicitable)

For any reasonable space of distributions (containing, e.g., all distributions with finite mean and a continuum of possible variances), there is no scoring function $S(x, y)$ such that $\text{CVaR}_q(F) = \arg\min_x \mathbb{E}_{Y \sim F}[S(x, Y)]$.

Implications. The non-elicitability of CVaR means there is no direct analogue of the simple violation-counting backtest for CVaR *alone* — one cannot construct a scoring function whose expected value, minimized over forecasts, recovers CVaR, and therefore cannot directly compare *realized losses* to a *single forecast number* $\widehat{\text{CVaR}}_q$ in a statistically principled way (in the sense of a "calibration" test analogous to the binomial test for VaR). This result, when first published, caused considerable consternation in the risk management community, since CVaR had by then become the preferred coherent alternative to VaR (Sections 4.1–4.3) precisely *because* of its superior theoretical properties — yet it appeared to lack a basic practical validation tool that VaR possessed.

Theorem (Fissler–Ziegel, 2016: joint elicitability of (VaR, CVaR))

The *pair* $(\text{VaR}_q(F), \text{CVaR}_q(F))$ is *jointly elicitable*: there exists a scoring function $S(x_1, x_2, y)$ (a function of two forecast values $x_1 \approx \text{VaR}_q$, $x_2 \approx \text{CVaR}_q$, and the realization y) such that $(\text{VaR}_q(F), \text{CVaR}_q(F)) = \arg\min_{(x_1, x_2)} \mathbb{E}_{Y \sim F}[S(x_1, x_2, Y)]$. An example of such a scoring function (one of a parametric family identified by Fissler–Ziegel) is

$$S(x_1, x_2, y) = (\mathbf{1}\{y \leq x_1\} - q)x_1 - \mathbf{1}\{y \leq x_1\}y + \frac{1}{1 - q}(x_2 - x_1)(\mathbf{1}\{y \leq x_1\} \cdot (\ldots)) + \ldots$$

(the full expression involves several terms; we omit the complete formula, which can be found in the original reference, but emphasize the qualitative conclusion).

Resolution and current practice. The joint elicitability result means: *report VaR and CVaR forecasts together*, and backtest the *pair* using a Fissler–Ziegel-type joint scoring function. This is now the basis of the standard CVaR backtesting procedure recommended in the Basel FRTB framework, which requires banks to report both $\widehat{\text{VaR}}_q$ and $\widehat{\text{CVaR}}_q$ and validates them jointly — often via a two-step procedure: first backtest VaR using the traditional binomial/Christoffersen tests (Section 4.5, first half), then conditionally backtest the CVaR forecast *given* that a VaR violation occurred (essentially, checking whether the *average* loss during violation episodes matches the forecast $\widehat{\text{CVaR}}_q$) — an approach sometimes called the "exceedance residual" test.

A broader lesson: the elicitability–coherence tension. This section's results illustrate a fundamental tension in risk measurement: the properties that make a risk measure *theoretically* desirable for decision-making and aggregation (coherence, Section 4.3 — which CVaR satisfies and VaR does not) are not the same as the properties that make it *statistically* validatable from data (elicitability — which VaR satisfies and CVaR, alone, does not). No risk measure satisfies both coherence and elicitability *except* in degenerate cases (Expectiles, a lesser-known risk measure based on asymmetric squared-error loss rather than the pinball loss, are both coherent and elicitable, but lack the direct "conditional tail expectation" interpretation of CVaR that makes it intuitively appealing — an active area of ongoing research and regulatory debate). This tension is not a flaw in any particular risk measure but reflects a genuine mathematical incompatibility, and awareness of it is essential for any practitioner choosing — and justifying the choice of — a risk measurement framework.

Appendix A

Supplementary Material: Extreme Value Theory

Additional theory, worked examples, and exercises extending Part 1.

Murphy diagrams and consistent scoring. Section 4.5 introduced the Fissler–Ziegel joint-elicitability result as the resolution of CVaR's non-elicitability. A complementary tool for *comparing* competing forecast systems — rather than evaluating a single system — is the *Murphy diagram*, which provides a visual representation of the consistent scoring functions for a given functional.

Definition (Murphy diagram for VaR)

For quantile forecasts $\hat{q}_t(q)$ and realizations Y_t $(t = 1, \dots, T)$, the *elementary scoring function* for the quantile functional at level q with "threshold" parameter θ is

$$S_\theta(x, y) = (\mathbf{1}\{y \le \theta\} - q)(\mathbf{1}\{x \le \theta\} - q), \qquad \theta \in \mathbb{R}.$$

The *Murphy diagram* plots the mean score difference between two competing forecasting systems F_1, F_2 as a function of θ:

$$\Delta(\theta) = \frac{1}{T} \sum_{t=1}^{T} S_\theta(F_{1,t}, Y_t) - S_\theta(F_{2,t}, Y_t),$$

for a range of θ values spanning the relevant loss range.

Interpretation. If $\Delta(\theta) > 0$ for *all* θ, then forecast system F_1 is uniformly worse than F_2 for VaR forecasting at level q, regardless of which consistent scoring function (which weighting of elementary scoring functions, Section 4.5's note on elicitability) is used. If $\Delta(\theta)$ changes sign across θ values, then neither system dominates: F_1 is better at some threshold levels and F_2 is better at others, and the preference depends on which region of the loss distribution one weights most heavily. **A Murphy diagram thus reveals whether one risk model is uniformly better or whether the comparison depends on the specific scoring function chosen** — a form of "dominance" analysis that goes beyond any single backtesting statistic (like Section 4.5's traffic-light violation count) by testing robustness to the choice of score.

Example — Numerical Murphy diagram comparison

Two VaR models (Model A: historical simulation; Model B: GPD-based EVT, as in Section 2.4) are compared over $T = 500$ days at $q = 0.99$. Suppose:

	Model A (hist. sim.)	Model B (EVT)
Violation count (should be ≈ 5)	8 (slight exceedance)	5 (on target)
Mean score at $\theta = \widehat{\text{VaR}}_{0.99}$	0.031	0.028

The Murphy diagram shows $\Delta(\theta) < 0$ (Model B better) for all $\theta > \$5\text{M}$ (the typical large-loss range), but $\Delta(\theta) > 0$ (Model A better) for $\theta < \$2\text{M}$ (the small-loss range). **Interpretation:** historical simulation (Model A) does a better job of modelling the "typical" or moderately bad day (small θ region, where most of the data lies), while the EVT model (Model B) is specifically better in the tail ($\theta > \$5\text{M}$), which is precisely the region relevant for risk management. A risk manager focused on tail events (a regulator, say, or a capital allocation decision-maker) should prefer Model B; a risk manager concerned with day-to-day P&L accuracy across the full distribution might

prefer Model A. The Murphy diagram makes this trade-off explicit and quantitative, in a way that neither the violation count alone (which also favours B at $q = 0.99$) nor a single mean-score comparison can.

Generalization to CVaR. An analogous Murphy diagram can be constructed for the VaR–CVaR pair, using the Fissler–Ziegel elementary scoring functions (Section 4.5), plotting the mean score difference as a function of a "threshold" parameter analogous to θ above but now for the joint functional. The resulting diagram provides a tool for comparing EVT-based versus parametric-tail-model versus historical-simulation CVaR forecasts across the full range of tail severity levels — the natural counterpart, for the joint-elicitable CVaR, of the VaR Murphy diagram developed here.

Closing remark: from measurement to management. The risk measures of Chapter 4 — VaR, CVaR, spectral measures — are statistical summaries of probability distributions. But risk *management*, as distinct from risk *measurement*, concerns actions: decisions about capital levels, portfolio composition, reinsurance purchasing, regulatory compliance. The connection between a number ($38.6M $\text{VaR}_{0.999}$, say, from Appendix G) and a decision (hold $38.6M in capital) involves a long chain of modelling assumptions (the GPD tail fit, the frequency model, the simulation), statistical uncertainties (parameter estimation error, Appendix G's sensitivity analysis), definitional choices (Section I.2's three return-period concepts), methodological choices (ML vs. L-moments, Section 2.4; which copula, Appendix J; which prior, Appendix K), and ultimately judgements about risk tolerance and conservatism that lie outside any model. **The mathematics of this chapter provides rigorous language and tools for the measurement step of this chain, but is neither designed nor able to substitute for the judgment required at the management step** — a distinction easily blurred when a complex analysis is summarized in a single reported number, and that this volume has tried, throughout, to keep in sharp focus.

A.1 The Delta Method for Extreme Value Functionals

Overview

Throughout Part 1, we quoted standard errors for GEV parameter estimates and derived quantities (return levels) without detailing the underlying delta-method machinery. This appendix section makes that machinery explicit and general, since it recurs throughout the volume whenever a nonlinear function of estimated parameters (a return level, a risk measure, a tail probability) requires a confidence interval.

Theorem (Multivariate delta method)

Let $\hat{\theta}_n \in \mathbb{R}^p$ be an estimator with $\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{d} \mathcal{N}(0, \Sigma)$, and let $g : \mathbb{R}^p \to \mathbb{R}$ be differentiable at θ with gradient $\nabla g(\theta) \neq 0$. Then

$$\sqrt{n}\left(g(\hat{\theta}_n) - g(\theta)\right) \xrightarrow{d} \mathcal{N}\left(0, \nabla g(\theta)^\top \Sigma \nabla g(\theta)\right).$$

Application to GEV-based quantiles. For $\theta = (\xi, \mu, \sigma)$ and $g(\theta) = x_p = \mu - \frac{\sigma}{\xi}[1-(-\log p)^{-\xi}]$ (the GEV quantile, Section 1.3), the gradient is

$$\frac{\partial x_p}{\partial \mu} = 1, \qquad \frac{\partial x_p}{\partial \sigma} = -\frac{1}{\xi}\left[1 - y_p^{-\xi}\right], \qquad \frac{\partial x_p}{\partial \xi} = \frac{\sigma}{\xi^2}\left[1 - y_p^{-\xi}\right] - \frac{\sigma}{\xi} y_p^{-\xi} \log y_p,$$

where $y_p = -\log p$. The asymptotic variance $\widehat{\mathrm{Var}}(\hat{x}_p) = \nabla g(\hat{\theta})^\top \hat{\Sigma} \nabla g(\hat{\theta})$ is what we computed numerically (in a GPD parametrization) in Section 2.5's worked example for the 100-year return level.

Example — Delta method for a GEV-based exceedance probability

Sometimes the quantity of interest is the *reverse* of a return level: given a fixed level x_0, what is the probability $p_0 = P(M > x_0) = 1 - G_{\xi,\mu,\sigma}(x_0)$ that a single block maximum exceeds x_0? Here

$$g(\theta) = 1 - \exp\left\{-\left[1 + \xi\frac{x_0 - \mu}{\sigma}\right]^{-1/\xi}\right\}.$$

Suppose $\hat{\xi} = 0.2, \hat{\mu} = 100, \hat{\sigma} = 20$, with $\hat{\Sigma} = \begin{pmatrix} 0.015 & -0.3 & 0.05 \\ -0.3 & 12 & -1 \\ 0.05 & -1 & 3 \end{pmatrix}$ (in the order (ξ, μ, σ)), and $x_0 = 180$.

Compute $z = 1 + \hat{\xi}(x_0 - \hat{\mu})/\hat{\sigma} = 1 + 0.2 \times 80/20 = 1 + 0.8 = 1.8$. Then $z^{-1/\hat{\xi}} = 1.8^{-5} \approx 0.0535$, and $g(\hat{\theta}) = 1 - \exp(-0.0535) \approx 0.0521$.

The gradient of g requires the chain rule through $\exp(-z^{-1/\xi})$; writing $h = z^{-1/\xi}$, $g = 1 - e^{-h}$, $\partial g/\partial \cdot = e^{-h} \times (-\partial h/\partial \cdot) = -e^{-h}\partial h/\partial \cdot$. With $e^{-h} \approx e^{-0.0535} \approx 0.9479$:

$$\frac{\partial h}{\partial \mu} = -\frac{1}{\xi} z^{-1/\xi - 1} \times \left(-\frac{\xi}{\sigma}\right) = \frac{z^{-1/\xi - 1}}{\sigma} = \frac{1.8^{-6}}{20} \approx \frac{0.0297}{20} \approx 0.00149.$$

$$\frac{\partial g}{\partial \mu} = -0.9479 \times 0.00149 \approx -0.00141.$$

Similarly for σ and ξ (more involved; we summarize): $\partial g/\partial \sigma \approx -0.00564$, $\partial g/\partial \xi \approx -0.0721$.

$$\mathrm{Var}(g(\hat{\theta})) \approx (-0.0721)^2(0.015) + (-0.00141)^2(12) + (-0.00564)^2(3) + \text{cross terms}.$$

$$\approx 0.0000780 + 0.0000239 + 0.0000954 + \text{cross terms} \approx 0.000197 + \text{cross terms}.$$

The cross terms (involving the off-diagonal covariances, each multiplied by $2\times$ the relevant pair of partial derivatives) typically partially cancel; suppose the total is $\mathrm{Var}(g(\hat{\theta})) \approx 0.00031$, giving SE ≈ 0.0176. A 95% CI for p_0 is $0.0521 \pm 1.96 \times 0.0176 \approx (0.0176, 0.0866)$ — the probability of exceeding $x_0 = 180$ in a given block could plausibly be anywhere from under 2% to nearly 9%, a more-than-fourfold range, again illustrating the substantial uncertainty inherent in extreme value extrapolation even for "moderate" extrapolation distances.

A.2 The r-largest Order Statistics Method

Overview

Block maxima (Section 1.5) use only the single largest observation per block, discarding potentially informative data on the second-, third-, etc. largest values. The *r-largest order statistics* method generalizes block maxima by modelling the joint distribution of the top r order statistics in each block, recovering additional statistical efficiency while retaining the GEV as the relevant limiting family.

Theorem (Joint limiting distribution of the r largest order statistics)

Let $X_{(n)} \geq X_{(n-1)} \geq \ldots \geq X_{(n-r+1)}$ denote the r largest of n i.i.d. observations from $F \in D(G_\xi)$. With the same normalizing constants a_n, b_n as in the FTG theorem,

$$\left(\frac{X_{(n)} - b_n}{a_n}, \ldots, \frac{X_{(n-r+1)} - b_n}{a_n} \right) \xrightarrow{d} (Z_1, \ldots, Z_r),$$

where $(Z_1, \ldots, Z_r)$ has joint density

$$f(z_1, \ldots, z_r) = g_\xi(z_r) \prod_{j=1}^{r} \left[1 + \xi z_j\right]^{-1/\xi - 1} \cdot \left[1 + \xi z_r\right]^{1/\xi}, \qquad z_1 > z_2 > \cdots > z_r,$$

expressible more cleanly as $f(z_1, \ldots, z_r) = \exp\left\{-(1+\xi z_r)^{-1/\xi}\right\} \prod_{j=1}^{r} (1+\xi z_j)^{-1/\xi - 1}$, for $1 + \xi z_j > 0$.

Likelihood and efficiency gain. Given m blocks, each contributing its top r order statistics, the log-likelihood for (ξ, μ, σ) sums the log of the joint density above (transformed back to the original scale) over all m blocks. Simulation studies show that, for $r = 2$ or $r = 3$, the asymptotic variance of $\hat{\xi}$ can be reduced by 20–40% relative to $r = 1$ (ordinary block maxima), with diminishing returns for larger r — and, importantly, with *increasing model risk*: the GEV-type joint limit for $(Z_1, \ldots, Z_r)$ is itself an approximation that degrades as r increases relative to the block size n (the lower order statistics "feel" the bulk of F more than the maximum does, so the asymptotic regime $n \to \infty$ with r fixed is reached more slowly for larger r).

Example — Numerical comparison: $r = 1$ vs $r = 3$

Suppose, for a dataset of $m = 40$ annual blocks (each of size $n = 365$), ordinary block-maxima ML gives $\hat{\xi}_{r=1} = 0.18$ with SE $= 0.11$. Applying the $r = 3$ likelihood to the same data (using the top 3 daily values from each year, $3 \times 40 = 120$ data points total, versus 40 for $r = 1$) might give $\hat{\xi}_{r=3} = 0.21$ with SE $= 0.075$ — a roughly 30% reduction in standard error, translating to noticeably tighter return-level confidence intervals (recall from Section 2.5's worked example how dominant the $\mathrm{Var}(\hat{\xi})$ term was in the return-level variance). The point estimates (0.18 vs 0.21) differ modestly; whether this reflects genuine information gain or mild bias from the $r = 3$ approximation being less accurate than the $r = 1$ approximation for this particular $n = 365$ would require further diagnostic checks (e.g. comparing $r = 1, 2, 3, 4, 5$ estimates for stability, analogous to the Hill-plot stability check of Section 2.3).

A.3 Worked Exercises for Part 1

Example — Exercise A.3.1 (Domain of attraction identification)

For each of the following survival functions, identify the domain of attraction (Gumbel, Fréchet, or Weibull) and, where applicable, the shape parameter ξ:

1. $\overline{F}(x) = (1+x)^{-3} e^{-x}$, $x \geq 0$.
2. $\overline{F}(x) = e^{-x^2}$, $x \geq 0$.
3. $\overline{F}(x) = (1-x)^2$, $0 \leq x \leq 1$.

4. $\overline{F}(x) = x^{-1}(\log x)^{-2}$, $x \geq e$.

Solution.

1. As $x \to \infty$, the exponential factor e^{-x} dominates the polynomial $(1+x)^{-3}$ — this is a (modified) exponential tail, placing F in the **Gumbel domain**, $\xi = 0$. (Verify via von Mises: $\overline{F}/f \to 1$ as $x \to \infty$, with derivative $\to 0$.)
2. $\overline{F}(x) = e^{-x^2}$ decays *faster* than any exponential — still Gumbel domain ($\xi = 0$), but with different normalizing constants ($a_n \sim (\log n)^{-1/2}$, even faster convergence than the ordinary exponential case).
3. $\overline{F}(x) = (1-x)^2$ on $[0,1]$: finite upper endpoint $x^* = 1$, with $\overline{F}(x^* - y) = \overline{F}(1-y) = y^2$ — polynomial approach to the endpoint with exponent 2. **Weibull domain**, $\xi = -1/2$ (since the exponent of $(x^* - x)$ in $\overline{F}$ corresponds to $\alpha = 1/|\xi|$, here $\alpha = 2 \Rightarrow \xi = -1/2$).
4. $\overline{F}(x) = x^{-1}(\log x)^{-2}$: regularly varying with index -1 (the $(\log x)^{-2}$ factor is slowly varying, Section 3.1). **Fréchet domain**, $\xi = 1/1 = 1$ — note $\xi = 1$ means even the *mean* of the limiting Fréchet distribution is infinite (Section 1.3's moment table: $k < 1/\xi = 1$ required for finite k-th moment, so even $k = 1$ fails at $\xi = 1$, technically the boundary case requiring $k < 1$ strictly).

Example — Exercise A.3.2 (GEV moment calculation)

For the GEV distribution with $\xi = 0.4, \mu = 0, \sigma = 1$, compute $\mathbb{E}[X]$, $\mathrm{Var}(X)$, and determine the highest finite moment.

Solution. Recall $\mathbb{E}[X] = (g_1 - 1)/\xi$ with $g_k = \Gamma(1 - k\xi)$, valid for $\xi < 1/k$.

$$g_1 = \Gamma(1 - 0.4) = \Gamma(0.6) \approx 1.4892, \qquad \mathbb{E}[X] = (1.4892 - 1)/0.4 \approx 1.223.$$

For the variance, need $g_2 = \Gamma(1 - 0.8) = \Gamma(0.2) \approx 4.5908$, valid since $\xi = 0.4 < 1/2 = 0.5$.

$$\mathrm{Var}(X) = (g_2 - g_1^2)/\xi^2 = (4.5908 - 1.4892^2)/0.16 = (4.5908 - 2.2177)/0.16 \approx 2.3731/0.16 \approx 14.83.$$

The highest finite moment: $k < 1/\xi = 2.5$, so $k = 1, 2$ are finite (mean, variance), but $k = 3$ requires $\xi < 1/3 \approx 0.333 < 0.4$ — **the third moment is infinite**. The distribution has finite mean and variance but undefined skewness.

Example — Exercise A.3.3 (Block size sensitivity)

A sequence of $N = 7300$ daily observations (20 years) is to be divided into blocks for GEV fitting. Compare the number of block maxima m and a rough assessment of bias/variance trade-off for block sizes $n = 30$ (monthly), $n = 91$ (quarterly), and $n = 365$ (annual).

Solution. $m = N/n$: monthly gives $m \approx 243$, quarterly $m \approx 80$, annual $m = 20$.

- Monthly ($m = 243$): abundant data for fitting (low variance), but $n = 30$ is small — the GEV approximation $F^{30} \approx G_\xi$ may carry substantial bias, particularly because monthly maxima of, e.g., daily temperature or rainfall, are strongly affected by seasonality (violating the i.i.d. assumption underlying FTG) — monthly blocks would need to be stratified by calendar month, effectively giving 12 separate analyses with $m \approx 20$ each, reducing to the annual case's sample size anyway.

- Quarterly ($m = 80$): intermediate; still likely seasonality issues unless stratified.
- Annual ($m = 20$): respects seasonality (each block spans a full cycle), $n = 365$ is plausibly large enough for reasonable GEV approximation, but $m = 20$ is a *very* small sample for estimating three parameters — standard errors on $\hat{\xi}$ will be large (cf. Section 1.5's worked example with $m = 50$, already showing $\text{SE}(\hat{\xi}) \approx 0.13$; $m = 20$ would be considerably worse, perhaps $\text{SE}(\hat{\xi}) \approx 0.20$ or more).

The seasonality consideration typically dominates in practice: annual blocking is standard for environmental data despite the resulting small m, motivating the POT/threshold methods of Chapter 2 (which can use all 7300 observations, with a declustering step to handle short-range dependence, rather than just $m = 20$ annual maxima) as the preferred approach whenever feasible.

Appendix B

Supplementary Material: Peaks-over-Threshold and the GPD

Additional theory, worked examples, and exercises extending Part 2.

B.1 Declustering and the Extremal Index

Overview

The peaks-over-threshold framework of Chapter 2 implicitly assumes *independent* exceedances. Real time series — river flows, temperatures, financial returns — exhibit *temporal dependence*: a large value today increases the probability of a large value tomorrow (e.g. a flood event spans several days). This section develops the *extremal index* $\theta \in (0, 1]$, which quantifies the degree of clustering of extremes, and the *declustering* procedures used to extract approximately independent "cluster maxima" for valid GPD/Poisson-process inference.

Definition (Extremal index)

For a stationary sequence $X_1, X_2, \ldots$ with marginal distribution F, the extremal index $\theta \in [0, 1]$ is defined by

$$\lim_{n\to\infty} P\left(\max(X_1, \ldots, X_n) \le u_n\right) = \left[\lim_{n\to\infty} F(u_n)^n\right]^\theta,$$

for sequences u_n such that $n(1 - F(u_n)) \to \tau \in (0, \infty)$.

Interpretation. $\theta = 1$ corresponds to (asymptotically) independent extremes — the i.i.d. theory of Chapters 1–2 applies directly. $\theta < 1$ indicates clustering: extremes tend to occur in groups, and the *effective* number of independent extreme events is reduced by a factor θ relative to the raw exceedance count. Equivalently, $1/\theta$ is (asymptotically) the mean cluster size — the average number of consecutive exceedances of a high threshold belonging to the same "episode."

Example — The extremal index for a simple AR(1)-driven process

Consider $X_t = \max(\rho X_{t-1}, \epsilon_t)$ for $\rho \in (0, 1)$ and ϵ_t i.i.d. with the same marginal distribution as X_t (a max-autoregressive process, a tractable model for clustered extremes). It can be shown that

$$\theta = 1 - \rho.$$

For $\rho = 0.7$ (moderately strong persistence), $\theta = 0.3$ — the mean cluster size is $1/\theta \approx 3.33$: an exceedance of a high threshold is, on average, followed by about 2.33 additional consecutive exceedances before the process drops below the threshold. If a raw POT analysis (Chapter 2) is applied *without* declustering to data from this process, the number of "exceedances" k used in GPD fitting (Section 2.3-2.4) overstates the effective sample size by a factor of $1/\theta \approx 3.33$ — standard errors on $\hat{\xi}, \hat{\beta}$ computed assuming i.i.d. exceedances would be too small by a factor of $\sqrt{3.33} \approx 1.83$, a substantial underestimation of uncertainty.

Declustering algorithms. The two standard approaches:

(*1*) *Blocks (runs) declustering.* Partition exceedance times into clusters separated by gaps of

at least r consecutive non-exceedances (the "run length"); take the maximum exceedance within each cluster as the cluster representative. The choice of r involves a bias-variance trade-off analogous to threshold and block-size choices elsewhere in this chapter: too small r leaves residual dependence between "declustered" maxima; too large r discards too many genuine, separate extreme events, merging them into artificially large clusters.

(2) *Ferro–Segers automatic declustering.* An automatic, data-driven method that estimates θ directly from the inter-exceedance times $T_1, T_2, \ldots$ (gaps between consecutive exceedance indices) via

$$\hat{\theta} = \frac{2\left(\sum_{i=1}^{k-1}(T_i - 1)\right)^2}{(k-1)\sum_{i=1}^{k-1}(T_i - 1)(T_i - 2)}$$

(for $\max T_i > 2$; a simpler formula applies if $\max T_i \leq 2$), and uses $\hat{\theta}$ directly to adjust the effective sample size and the Poisson rate λ of Section 2.5, without requiring an arbitrary run-length parameter r.

B.2 The Poisson–GPD Model in Full

Overview

Section 2.5 introduced the Poisson process for exceedance *times* somewhat informally. Here we present the full Poisson–GPD model as a point process on (time) $\times$ (exceedance magnitude), derive its likelihood, and show its exact equivalence to the GEV model for block maxima under the appropriate parameter correspondence — making precise the connection alluded to throughout Chapter 2.

Definition (Poisson–GPD point process model)

Exceedances of threshold u occur as a Poisson process in time with rate λ (exceedances per unit time), and each exceedance has magnitude $Y_i = X_i - u \sim H_{\xi,\beta}$ (GPD) independently. Equivalently, the points $\{(t_i, X_i) : X_i > u\}$ form a non-homogeneous Poisson process on $[0, T] \times (u, \infty)$ with intensity measure

$$\Lambda(dt \times dx) = \lambda \left[1 + \xi\frac{x-u}{\beta}\right]_+^{-1/\xi - 1} \frac{1}{\beta}\, dt\, dx,$$

i.e. rate λ in time, density $h_{\xi,\beta}(x-u)$ in magnitude.

Equivalence to GEV. The annual maximum $M = \max\{X_i : t_i \in [0,1]\}$ (taking the unit of time as one year) under this point process model has $P(M \leq x) = P(\text{no points in } [0,1] \times (x, \infty)) = \exp(-\Lambda([0,1]\times(x,\infty)))$ for $x > u$. Computing $\Lambda([0,1]\times(x,\infty)) = \lambda\left[1 + \xi\frac{x-u}{\beta}\right]^{-1/\xi}$ (for $x > u$), so

$$P(M \leq x) = \exp\left\{-\lambda\left[1 + \xi\frac{x-u}{\beta}\right]^{-1/\xi}\right\}.$$

Comparing to the GEV CDF $G_{\xi,\mu,\sigma}(x) = \exp\{-[1 + \xi(x-\mu)/\sigma]^{-1/\xi}\}$, we identify

$$\sigma = \beta\lambda^{\xi}, \qquad \mu = u - \frac{\beta}{\xi}(1 - \lambda^{\xi}) = u + \frac{\sigma - \beta}{\xi}.$$

This is an exact equivalence: the GEV parameters (μ, σ) are deterministic functions of the POT parameters (u, λ, ξ, β) (with ξ shared), confirming PBdH's prediction (Section 2.2) at

the level of the full models, not merely asymptotically.

Example — Numerical translation between POT and GEV parametrizations

Using the running example's POT parameters $u = 30, \hat{\xi} = 0.34, \hat{\beta} = 9.8, \hat{\lambda} = 4.5$:

$$\hat{\sigma}_{\text{GEV}} = 9.8 \times 4.5^{0.34}.$$

Compute $4.5^{0.34} = \exp(0.34 \times \ln 4.5) = \exp(0.34 \times 1.504) = \exp(0.5114) \approx 1.668$.

$$\hat{\sigma}_{\text{GEV}} \approx 9.8 \times 1.668 \approx 16.34.$$

$$\hat{\mu}_{\text{GEV}} = 30 + \frac{16.34 - 9.8}{0.34} = 30 + \frac{6.54}{0.34} \approx 30 + 19.24 \approx 49.24.$$

So the equivalent annual-maximum GEV model has $(\hat{\xi}, \hat{\mu}, \hat{\sigma}) \approx (0.34, 49.24, 16.34)$. One can verify this is consistent: the 100-year return level under this GEV (Section 1.3's formula) should match the $x_{100} \approx 231.1$ computed in Section 2.5's POT-based calculation. Using the GEV quantile formula with $p = 0.99$: $(-\log 0.99)^{-\xi} = (0.01005)^{-0.34} \approx \exp(0.34 \times 4.600) = \exp(1.564) \approx 4.778$.

$$x_{0.99} = 49.24 - \frac{16.34}{0.34}(1 - 4.778) = 49.24 - 48.06 \times (-3.778) \approx 49.24 + 181.6 \approx 230.8,$$

matching the POT-based 231.1 up to rounding error — confirming the exact parameter correspondence.

B.3 Worked Exercises for Part 2

Example — Exercise B.3.1 (GPD likelihood for $\xi = 0$)

Derive the MLE for β in the special case $\xi = 0$ (Exponential exceedances), and verify it coincides with the method-of-moments estimator.

Solution. For $\xi = 0$, $H_{0,\beta}(y) = 1 - e^{-y/\beta}$, density $h(y) = \frac{1}{\beta} e^{-y/\beta}$. Log-likelihood for $Y_1, \dots, Y_k$ i.i.d.:

$$\ell(\beta) = -k \log \beta - \frac{1}{\beta} \sum_i Y_i.$$

$\partial \ell / \partial \beta = -k/\beta + \frac{1}{\beta^2} \sum Y_i = 0 \Rightarrow \hat{\beta} = \frac{1}{k} \sum Y_i = \bar{Y}$, the sample mean — which is also the method-of-moments estimator ($\mathbb{E}[Y] = \beta$ for Exponential), as expected.

Example — Exercise B.3.2 (Threshold stability check)

Given exceedances of threshold $u_1 = 20$ fitted as GPD($\xi = 0.25, \beta_1 = 8$), predict the GPD parameters for threshold $u_2 = 35$, and verify using the threshold-stability formula of Section 2.1.

Solution. By threshold stability, excesses over $u_2 = 35$ (i.e. excesses over $u_1 = 20$ of an additional $v = u_2 - u_1 = 15$) follow GPD$(\xi, \beta_1 + \xi v)$ = GPD$(0.25, 8 + 0.25 \times 15)$ =GPD$(0.25, 8 + 3.75)$ =GPD$(0.25, 11.75)$.

Verification via direct survival function. $\overline{H}_{0.25,8}(15 + y)/\overline{H}_{0.25,8}(15) \stackrel{?}{=} \overline{H}_{0.25,11.75}(y)$.

LHS $= \frac{(1+0.25(15+y)/8)^{-4}}{(1+0.25\times 15/8)^{-4}} = \left(\frac{1+0.25(15+y)/8}{1+0.25\times 15/8}\right)^{-4} = \left(\frac{8+0.25(15+y)}{8+0.25\times 15}\right)^{-4} = \left(\frac{11.75+0.25y}{11.75}\right)^{-4} =$

$\left(1+\frac{0.25y}{11.75}\right)^{-4}$. RHS $=(1+0.25y/11.75)^{-1/0.25}=(1+0.25y/11.75)^{-4}$. LHS = RHS. ✓

Example — Exercise B.3.3 (Effective sample size with clustering)

A POT analysis identifies $k=300$ raw exceedances of a threshold over 30 years of daily data. The Ferro–Segers estimator gives $\hat{\theta}=0.4$. (a) Estimate the effective number of independent clusters. (b) How does this affect the Poisson rate λ used in return-level calculations (Section 2.5)?

Solution. (a) Effective independent clusters $\approx k\times\hat{\theta}=300\times0.4=120$. (b) The *raw* exceedance rate is $300/30=10$/year. For return-level calculations, the relevant rate is the rate of *cluster* (episode) occurrences, $\lambda_{\text{eff}}=\hat{\theta}\times\lambda_{\text{raw}}=0.4\times10=4$/year — using the raw rate $\lambda=10$ in the return-level formula of Section 2.5 would *overstate* the frequency of independent extreme episodes by a factor of $1/\hat{\theta}=2.5$, leading to return-level estimates that are systematically too *low* for a given return period T (since higher λ at fixed $\overline{H}_{\xi,\beta}(x_T-u)=1/(\lambda T)$ requires smaller $\overline{H}$, i.e. smaller x_T-u). Both λ and the GPD parameters (ξ,β) should, strictly, be estimated from declustered (cluster-maximum) data for full internal consistency, though in practice (ξ,β) are sometimes estimated from all raw exceedances (since cluster maxima and raw exceedances above a high threshold often have similar marginal distributions, even though their *joint*/temporal structure differs) while only λ is adjusted by $\hat{\theta}$.

Appendix C

Supplementary Material: Heavy Tails, Power Laws, and Multivariate Extremes

Additional theory, worked examples, and exercises extending Part 3.

C.1 Second-Order Regular Variation and Bias-Corrected Estimators

Overview

Section 1.4 mentioned second-order regular variation as governing convergence *rates*. Here we make this precise and show how it motivates bias-corrected variants of the Hill estimator (Section 2.3) — estimators that remain consistent and asymptotically normal while reducing the small-sample bias that plagues the ordinary Hill estimator at larger k.

Definition (Second-order regular variation)

$\overline{F} \in 2RV(\alpha, \rho)$ if $\overline{F}(x) = cx^{-\alpha}(1 + A(x) + o(A(x)))$ as $x \to \infty$, where $A(x) \to 0$ and $A \in RV_\rho$ for some $\rho \leq 0$ (the *second-order parameter*).

Effect on the Hill estimator. Under $2RV(\alpha, \rho)$, the Hill estimator $\hat{\xi}_k$ (Section 2.3) has an asymptotic bias of order $A(n/k)$, in addition to its variance ξ^2/k. The mean squared error

$$\text{MSE}(\hat{\xi}_k) \approx \frac{\xi^2}{k} + [c_A \cdot A(n/k)]^2$$

is minimized at some intermediate k^* that balances these two terms — the theoretical underpinning of the bias-variance trade-off in Hill-plot interpretation (Section 2.3), now made quantitative. The optimal k^* scales as $k^* \sim n^{-2\rho/(1-2\rho)}$ (for $\rho < 0$), and $\text{MSE}(\hat{\xi}_{k^*}) \sim n^{2\rho/(1-2\rho)}$ — both vanishing as $n \to \infty$, but at a rate slower than the "parametric" rate n^{-1} unless $\rho \to -\infty$ (i.e. unless the second-order term is negligible, the "ideal" Pareto case).

Definition (Bias-corrected Hill estimator, simplified)

A common bias correction replaces $\hat{\xi}_k$ with

$$\hat{\xi}_k^{\text{BC}} = \hat{\xi}_k \left(1 - \frac{\hat{\rho}}{1 - \hat{\rho}} \left(\frac{\hat{\xi}_k}{\hat{\xi}_k^{(2)}} - 1\right)\right)$$

or similar formulas involving an auxiliary estimator $\hat{\rho}$ of the second-order parameter and a secondary statistic $\hat{\xi}_k^{(2)}$ (often based on a different, typically larger, number of order statistics); the exact formulas vary across the several competing bias-correction proposals in the literature (de Haan–Peng, Gomes et al., Caeiro–Gomes, among others), all sharing the goal of extending the "stable region" of the Hill plot to larger k, thereby reducing variance without incurring the bias that would afflict the uncorrected estimator at that k.

Example — Illustrating bias correction

Suppose true $\overline{F}(x) = x^{-2}(1+2x^{-1})$ for large x — i.e. $\alpha = 2$ ($\xi = 0.5$) with second-order parameter $\rho = -1$ (since $A(x) = 2x^{-1} \in RV_{-1}$). For $n = 5000$:

Uncorrected Hill, various k:

k	$\mathbb{E}[\hat{\xi}_k]$ (approx.)	SD$(\hat{\xi}_k)$
50	0.52	0.071
200	0.56	0.035
500	0.63	0.022
1000	0.71	0.016

The bias grows with k (from $+0.02$ at $k = 50$ to $+0.21$ at $k = 1000$) while the SD shrinks — the classic trade-off. The bias-corrected estimator, at $k = 1000$, might achieve $\mathbb{E}[\hat{\xi}_k^{\text{BC}}] \approx 0.51$ with SD ≈ 0.017 — nearly eliminating the bias while retaining most of the variance reduction from using $k = 1000$ rather than $k = 50$. This roughly $4\times$ reduction in MSE (from $(0.21)^2 + 0.016^2 \approx 0.0444$ to $(0.01)^2 + 0.017^2 \approx 0.00039$, over $100\times$ smaller, though these illustrative numbers somewhat exaggerate the typical gain) illustrates why bias correction is standard practice in serious applied tail-index estimation, despite the additional complexity of estimating ρ.

C.2 Extremes of Dependent Sequences: Beyond i.i.d.

Overview

Appendix B.1 introduced the extremal index θ for clustered extremes in a time-series context. This section broadens the perspective: many real-world "extreme" phenomena arise not from a single time series but from *spatial* fields (e.g. rainfall over a region) or from *multivariate* systems with complex dependence (Section 3.5). We outline max-stable processes — the spatial analogue of the GEV — as the natural extension.

Max-stable processes. A stochastic process $\{Z(s) : s \in S\}$ (e.g. $S \subset \mathbb{R}^2$, a spatial domain) is *max-stable* if, for i.i.d. copies $Z_1, Z_2, \ldots$ of the process and any n, there exist (deterministic) functions $a_n(s) > 0, b_n(s)$ such that

$$\left\{ \frac{\max_{i \le n} Z_i(s) - b_n(s)}{a_n(s)} \right\}_{s \in S} \overset{d}{=} \{Z(s)\}_{s \in S}.$$

Just as the GEV (Section 1.3) is the unique possible limit for normalized maxima of a *single* random variable, max-stable processes are the unique possible limits for normalized *pointwise* maxima of a spatial field — and the marginal distribution at each s is necessarily GEV.

The Smith and Schlather models. Two widely used parametric families of max-stable processes:

Smith model: $Z(s) = \max_i \zeta_i \, \phi(s - U_i; \Sigma)$, where $\{(\zeta_i, U_i)\}$ is a Poisson process on $(0, \infty) \times \mathbb{R}^2$ with intensity $\zeta^{-2} d\zeta \, dU$, and $\phi(\cdot; \Sigma)$ is a Gaussian density with covariance Σ — Σ controls the spatial range of dependence.

Schlather model: $Z(s) = \max_i \zeta_i \max(0, Y_i(s))$ for a Gaussian process Y_i with correlation function $\rho(\cdot)$ — often more flexible and computationally tractable than Smith.

Pairwise extremal dependence. For both models, the bivariate dependence between $Z(s_1)$ and $Z(s_2)$ is governed by a function of the distance $\|s_1 - s_2\|$, analogous to the tail dependence coefficient λ_U of Section 3.5 but now varying continuously with spatial separation — $\lambda_U(\|s_1 - s_2\|) \to 1$ as $\|s_1 - s_2\| \to 0$ (nearby locations are perfectly tail-dependent) and $\lambda_U(\|s_1 - s_2\|) \to 0$ as $\|s_1 - s_2\| \to \infty$ (distant locations become tail-independent), with the rate of decay determined by Σ (Smith) or $\rho(\cdot)$ (Schlather).

Example — Practical relevance: regional flood/heatwave risk

For a power grid or insurance portfolio exposed to weather-driven risk across a region, the spatial extremal dependence directly determines aggregate risk: if extreme heatwaves (driving simultaneous air-conditioning demand spikes) are spatially tail-*dependent* across a service territory (as is physically realistic — heatwaves are large-scale meteorological phenomena), the aggregate peak demand cannot be modelled by simply summing independent location-level GEV/GPD models (which would substantially *underestimate* the probability of a region-wide simultaneous extreme, by the same logic as Section 3.5's Gaussian-copula warning). Max-stable process models, fitted to historical spatial temperature fields, provide the correct framework for such aggregate risk assessment — a topic developed further in the network-resilience context of Chapter 14 (Volume I, Part IV) and the systemic risk context of Chapter 16.

C.3 Worked Exercises for Part 3

Example — Exercise C.3.1 (Regular variation arithmetic)

If $f \in RV_2$ and $g \in RV_{-3}$, determine the regular variation index (if any) of: (a) fg; (b) f/g; (c) $f+g$; (d) $f \circ g$ (composition, where meaningful).

Solution. (a) $fg \in RV_{2+(-3)} = RV_{-1}$ (indices of regularly varying functions add under multiplication, since $L_1(x)L_2(x)$ is again slowly varying if L_1, L_2 are). (b) $f/g \in RV_{2-(-3)} = RV_5$. (c) $f+g$: since $2 > -3$, f dominates for large x ($f(x)/g(x) \to \infty$), so $f+g \sim f \in RV_2$ (the sum is regularly varying with the index of the *larger*-index summand, generically). (d) $f \circ g$: requires $g(x) \to \infty$ as $x \to \infty$ for $f \circ g$ to be evaluated in the regime where f's regular variation applies; but $g \in RV_{-3}$ means $g(x) \to 0$ as $x \to \infty$ — so $f(g(x))$ involves f evaluated near 0, not at infinity, and the regular variation of f at infinity does not directly apply. The composition is generally not regularly varying at infinity under these conditions without further information about f near 0.

Example — Exercise C.3.2 (Subexponentiality check)

Determine whether the Weibull distribution with shape parameter τ, $\overline{F}(x) = e^{-x^\tau}$, is subexponential, for (a) $\tau = 0.5$ and (b) $\tau = 2$.

Solution. The Weibull distribution with $\overline{F}(x) = e^{-x^\tau}$ is subexponential if and only if $\tau < 1$ (the case $\tau = 1$ is the boundary Exponential, also subexponential by inclusion in $RV_{-\infty}$-adjacent reasoning, but more directly: Exponential itself is *not* subexponential in the strict sense of Section 3.3's definition — recheck: actually the Exponential distribution fails the subexponential ratio condition, as discussed in Section 3.3's remarks). For $\tau < 1$: $\overline{F}(x)e^{\epsilon x} = e^{\epsilon x - x^\tau} \to \infty$ for any $\epsilon > 0$ (since x eventually dominates x^τ for $\tau < 1$), confirming the tail decays slower than exponential — the defining heuristic for subexponentiality, and indeed $\tau < 1$ Weibull is a standard example of a subexponential, non-regularly-varying distribution. (a) $\tau = 0.5 < 1$: **subexponential**. (b) $\tau = 2 > 1$: $\overline{F}(x) = e^{-x^2}$ decays *faster* than exponential — **not subexponential** (it is in fact in the Gumbel domain with extremely light tails, as noted in Exercise A.3.1(ii)).

Example — Exercise C.3.3 (Tail dependence from a mixture model)

Let (X, Y) be generated as follows: with probability p, $X = Y = Z$ for $Z \sim$Pareto(α); with probability $1-p$, X, Y are independent draws from Pareto(α). Compute λ_U as a function of p.

Solution. For large threshold u (with $\overline{F}(u) = u^{-\alpha} \to 0$): $P(X > u, Y > u) = p\,P(Z > u) + (1-p)P(X > u)P(Y > u) = p\,u^{-\alpha} + (1-p)u^{-2\alpha}$. As $u \to \infty$, the second term is $o(u^{-\alpha})$ (since $2\alpha > \alpha$), so $P(X > u, Y > u) \sim p\,u^{-\alpha}$. Also $P(X > u) \sim u^{-\alpha}$ (the mixture doesn't change the marginal tail to leading order, since both components contribute $O(u^{-\alpha})$ marginally — more precisely $P(X > u) = p\,P(Z > u) + (1-p)P(X > u$ in indep. part$) = pu^{-\alpha} + (1-p)u^{-\alpha} = u^{-\alpha}$, exactly, for this construction). So

$$\lambda_U = \lim_{u \to \infty} \frac{P(X > u, Y > u)}{P(X > u)} = \frac{p\,u^{-\alpha}}{u^{-\alpha}} = p.$$

The tail dependence coefficient equals the mixing probability p — an intuitive and clean result: p is exactly the probability that an extreme event is of the "shared/systemic" type (where X and Y necessarily move together) rather than the "idiosyncratic/independent" type. This simple mixture model is a useful building block for constructing tractable examples of arbitrary tail dependence $\lambda_U \in [0, 1]$, and underlies some of the simplified systemic-risk models discussed in Chapter 16.

Appendix D

Supplementary Material: Risk Measures, Estimation, and Synthesis

Additional theory, worked examples, exercises, and a synthesis connecting Volume I to the rest of the series.

D.1 Estimating Risk Measures Under Parameter Uncertainty: A Bayesian Treatment

Overview

Sections 2.4–2.5 and 4.1–4.2 treated GPD/GEV parameters and derived risk measures (VaR, CVaR) via point estimation (ML, PWM) with delta-method confidence intervals. This appendix develops the Bayesian alternative in more depth, addressing a conceptual point of real practical importance: the *predictive* distribution for a future extreme observation differs from — and has heavier tails than — the fitted GPD/GEV evaluated at point-estimated parameters, because it integrates over parameter uncertainty.

Definition (Posterior predictive distribution)

Given data $\mathbf{y}$ and prior $\pi(\theta)$ for GPD parameters $\theta = (\xi, \beta)$, the posterior is $\pi(\theta \mid \mathbf{y}) \propto L(\theta; \mathbf{y})\pi(\theta)$. The posterior predictive distribution for a new excess Y^* is

$$P(Y^* > y \mid \mathbf{y}) = \int \overline{H}_{\xi,\beta}(y)\, \pi(\theta \mid \mathbf{y})\, d\theta.$$

Why the predictive distribution is heavier-tailed. Even if the posterior is concentrated around $\hat{\xi} \approx 0.3$, the predictive survival function $P(Y^* > y \mid \mathbf{y})$ is an *average* of $\overline{H}_{\xi,\beta}(y)$ over the posterior, including the posterior's upper tail where ξ may be substantially larger than 0.3. Since $\overline{H}_{\xi,\beta}(y)$ for large y is extremely sensitive to ξ (recall $\overline{H}_{\xi,\beta}(y) \sim (\xi y/\beta)^{-1/\xi}$, and $1/\xi$ appears in the *exponent*), this averaging is dominated, for large enough y, by the largest plausible values of ξ in the posterior — a manifestation of Jensen's inequality applied to a highly convex function of ξ.

Example — Numerical illustration: plug-in vs. predictive tail probability

Suppose the posterior for ξ (given β fixed at $\hat{\beta} = 10$ for simplicity) is approximately $\mathcal{N}(0.3, 0.01)$ (i.e. posterior mean 0.3, SD 0.1, truncated to $\xi > 0$). Compute $P(Y^* > 100)$ two ways:

Plug-in (point estimate): $\overline{H}_{0.3,10}(100) = (1 + 0.3 \times 100/10)^{-1/0.3} = (1+3)^{-3.333} = 4^{-3.333}$. Compute $4^{-3.333} = \exp(-3.333 \times \ln 4) = \exp(-3.333 \times 1.386) = \exp(-4.621) \approx 0.00982$.

Predictive (averaging over $\xi \sim \mathcal{N}(0.3, 0.01)$, via numerical integration / simulation): sample $\xi^{(i)} \sim \mathcal{N}(0.3, 0.01)$, compute $\overline{H}_{\xi^{(i)},10}(100)$ for each, average. Because $\overline{H}_{\xi,10}(100)$ increases sharply in ξ (e.g. at $\xi = 0.4$: $(1+4)^{-2.5} = 5^{-2.5} = \exp(-2.5 \times 1.609) = \exp(-4.024) \approx 0.0179$, nearly double the $\xi = 0.3$ value; at $\xi = 0.5$: $(1+5)^{-2} = 6^{-2} \approx 0.0278$, nearly triple), the average over $\xi \sim \mathcal{N}(0.3, 0.1)$ is pulled well above the plug-in value at $\xi = 0.3$. A numerical average might give ≈ 0.0135 — roughly 37% higher than the plug-in estimate of 0.00982.

Practical consequence. A risk manager using the plug-in (point-estimate) GPD sys-

tematically *understates* the probability of large exceedances — and correspondingly understates VaR and CVaR at extreme quantiles — relative to a fully Bayesian (or otherwise uncertainty-propagating) assessment, even when the point estimate $\hat{\xi}$ itself is unbiased. This is sometimes called *estimation risk* or *parameter uncertainty risk*, distinct from the "model risk" of choosing the wrong distributional family altogether, and it is a systematic, one-directional bias (always toward *understating* tail risk for convex tail functionals) — not a symmetric "noise" that averages out.

D.2 Risk Measure Aggregation and the Subadditivity–Diversification Distinction

Overview

Sections 3.3 and 4.3 separately discussed (a) the failure of diversification benefit for heavy-tailed *sums* (single-big-jump) and (b) the subadditivity axiom for coherent risk measures. This appendix clarifies the relationship — and non-equivalence — between these two concepts, a frequent source of confusion.

Subadditivity is about the risk *measure*; diversification benefit is about the risk *measure applied to a specific aggregation*. ρ subadditive means $\rho(L_1 + L_2) \leq \rho(L_1) + \rho(L_2)$ *for all* L_1, L_2 — a universal property of ρ. "Diversification benefit" for a *specific* pair (L_1, L_2) means the inequality is *strict* (and, ideally, by a substantial margin): $\rho(L_1 + L_2) \ll \rho(L_1) + \rho(L_2)$.

A subadditive (even coherent) ρ can yield *near-equality* $\rho(L_1+L_2) \approx \rho(L_1)+\rho(L_2)$ for specific (L_1, L_2) — e.g. i.i.d. heavy-tailed L_1, L_2 with CVaR as in Section 4.2's worked example, which achieved *exact equality*. Subadditivity guarantees $\rho(L_1 + L_2)$ is *never worse* than the sum, but says nothing about *how much better* it might be — and for heavy-tailed, tail-dependent risks (Section 3.5), "how much better" can be arbitrarily close to "not at all," even for a fully coherent ρ.

Example — Quantifying diversification benefit: the diversification ratio

Define $DR(L_1, L_2) = \dfrac{\rho(L_1) + \rho(L_2)}{\rho(L_1 + L_2)} \geq 1$ (for coherent ρ, by subadditivity). $DR = 1$ means no diversification benefit; $DR \to \infty$ would mean (in principle) unlimited benefit, though in practice DR is bounded.

For two i.i.d. light-tailed (Gaussian) risks with correlation ρ_{corr}, using CVaR_q (which for Gaussians is proportional to the standard deviation): $\rho(L_1 + L_2) \propto \sqrt{2 + 2\rho_{\text{corr}}}\,\sigma$ while $\rho(L_1) + \rho(L_2) = 2\sigma$, giving $DR = \frac{2}{\sqrt{2+2\rho_{\text{corr}}}} = \sqrt{\frac{2}{1+\rho_{\text{corr}}}}$. For $\rho_{\text{corr}} = 0$: $DR = \sqrt{2} \approx 1.414$ — substantial benefit. For $\rho_{\text{corr}} = 1$ (perfect correlation): $DR = 1$ — no benefit, as expected (perfectly correlated risks don't diversify).

For two i.i.d. *heavy-tailed* (Fréchet-domain, $\xi > 0$) risks, even with $\rho_{\text{corr}} = 0$ (independence), the single-big-jump principle (Section 3.3) implies $\text{CVaR}_q(L_1 + L_2) \to 2\,\text{CVaR}_q(L_1)$ as $q \to 1$ (the confidence level approaches the extreme tail) — i.e. $DR \to 1$ even under *independence*, for sufficiently extreme q. **The diversification ratio for heavy-tailed risks degrades toward 1 as one moves further into the tail**, regardless of correlation — a quantitative restatement of Section 3.3's message, now expressed in risk-measure terms and showing precisely how "coherent but no diversification benefit" (independence, $DR \to 1$) differs from "not coherent, diversification

appears actively harmful" (VaR's $DR < 1$, Section 4.1's counterexample).

D.3 Worked Exercises for Part 4

Example — Exercise D.3.1 (CVaR for a normal loss)

For $L \sim \mathcal{N}(\mu, \sigma^2)$, derive $\text{CVaR}_q(L)$ in closed form.
Solution. $\text{VaR}_q(L) = \mu + \sigma\Phi^{-1}(q)$. The CVaR formula $\text{CVaR}_q(L) = \mathbb{E}[L \mid L \geq \text{VaR}_q(L)]$. For $Z = \frac{L-\mu}{\sigma} \sim \mathcal{N}(0,1)$, $\mathbb{E}[Z \mid Z \geq z_q] = \frac{\phi(z_q)}{1-q}$ where $z_q = \Phi^{-1}(q)$ and ϕ is the standard normal density (this is the standard "inverse Mills ratio" result). So

$$\text{CVaR}_q(L) = \mu + \sigma\frac{\phi(\Phi^{-1}(q))}{1-q}.$$

For $q = 0.99$: $\Phi^{-1}(0.99) \approx 2.326$, $\phi(2.326) = \frac{1}{\sqrt{2\pi}}e^{-2.326^2/2} \approx 0.0267$. So $\text{CVaR}_{0.99}(L) \approx \mu + \sigma \times \frac{0.0267}{0.01} = \mu + 2.67\sigma$, compared to $\text{VaR}_{0.99}(L) \approx \mu + 2.326\sigma$ — CVaR exceeds VaR by about 0.34σ for the Gaussian case, a much smaller gap (relative to σ) than the heavy-tailed example of Section 4.2 (where the gap was over 50% of VaR itself), illustrating again how the CVaR-VaR gap is itself an informative tail-heaviness diagnostic.

Example — Exercise D.3.2 (Coherence verification for a proposed risk measure)

Consider $\rho(L) = \mathbb{E}[L] + c \cdot \text{SD}(L)$ for constant $c > 0$ (the "mean-plus-c-sigma" rule, common in engineering safety margins). Determine which coherence axioms it satisfies and which it violates.
Solution. *Translation invariance:* $\rho(L+k) = \mathbb{E}[L+k] + c\,\text{SD}(L+k) = \mathbb{E}[L] + k + c\,\text{SD}(L) = \rho(L) + k$. **Satisfied.** *Positive homogeneity:* $\rho(\lambda L) = \lambda\mathbb{E}[L] + c\lambda\,\text{SD}(L) = \lambda\rho(L)$ for $\lambda > 0$. **Satisfied.** *Subadditivity:* $\rho(L_1 + L_2) = \mathbb{E}[L_1] + \mathbb{E}[L_2] + c\,\text{SD}(L_1 + L_2)$. Since $\text{SD}(L_1 + L_2) = \sqrt{\text{Var}(L_1) + \text{Var}(L_2) + 2\text{Cov}(L_1, L_2)} \leq \text{SD}(L_1) + \text{SD}(L_2)$ (triangle inequality for the L^2 norm), $\rho(L_1 + L_2) \leq \rho(L_1) + \rho(L_2)$. **Satisfied**, with equality iff L_1, L_2 are perfectly positively correlated. *Monotonicity:* Here is where it **fails**. Consider $L_1 \equiv 5$ (deterministic) and $L_2 \sim \mathcal{N}(5, 1)$ (so $L_2 \geq L_1$ is *not* guaranteed, but suppose instead we construct $L_1 \leq L_2$ a.s. by, e.g., $L_2 = L_1 + |\mathcal{N}(0,1)|$, so $L_2 \geq L_1 = 5$ always). Then $\mathbb{E}[L_1] = 5, \text{SD}(L_1) = 0 \Rightarrow \rho(L_1) = 5$. $\mathbb{E}[L_2] = 5 + \mathbb{E}|\mathcal{N}(0,1)| = 5 + \sqrt{2/\pi} \approx 5.798$, $\text{SD}(L_2) = \text{SD}(|\mathcal{N}(0,1)|) = \sqrt{1 - 2/\pi} \approx 0.603$. For $c = 1$: $\rho(L_2) \approx 5.798 + 0.603 = 6.401 > \rho(L_1) = 5$. In this case monotonicity *holds*. However, one can construct examples (with $L_1 \leq L_2$ a.s. but $\text{SD}(L_1) > \text{SD}(L_2)$ by a sufficient margin — e.g. L_1 highly variable but with low mean, $L_2 = L_1 + \delta$ for small constant δ shifting up slightly while... actually translation by a constant preserves SD) — more carefully: take $L_1 \sim \text{Uniform}(0, 10)$ (mean 5, SD ≈ 2.89) and $L_2 \equiv 5.5$ (deterministic, mean 5.5, SD 0), with *no* a.s. ordering possible since L_1 ranges over $(0, 10) \not\subseteq \{5.5\}$ – monotonicity requires $L_1 \leq L_2$ a.s., impossible here. The mean-plus-sigma rule's monotonicity failures are subtle and require care to exhibit; nonetheless, it is a known fact that $\rho(L) = \mathbb{E}[L] + c\,\text{SD}(L)$ is **not monotonic in general** for $c > 0$ when L can take negative values or when comparing distributions of different shapes, because SD does not respect the a.s. ordering (a higher-variance variable can be stochastically *smaller*). This is the reason mean-variance-based risk measures, despite their familiarity from portfolio theory, are not coherent and are increasingly supplemented or replaced by CVaR-based measures in tail-risk-focused applications.

D.4 Synthesis: Extreme Value Theory as the Foundation for Volume I

This closing section synthesizes the chapter's results and previews their role in the remainder of Volume I and the wider *Mathematics of Robustness, Risk & Failure* series.

The unifying thread. Chapter 1 established that extremes of i.i.d. sequences converge to one of three universal types (GEV, Section 1.2–1.3), classified by the tail behaviour of the underlying distribution (domains of attraction, Section 1.4) and estimable from data via block maxima (Section 1.5). Chapter 2 showed the dual perspective — excesses over thresholds converge to the GPD (PBdH, Section 2.2), sharing the same shape parameter ξ, with practical threshold-selection (Section 2.3) and estimation (Section 2.4) methodology, culminating in return-level/return-period calculations (Section 2.5) that directly translate extreme value models into actionable risk statements. Chapter 3 generalized the tail-behaviour classification via regular variation (Section 3.1), confronted the empirical and statistical pitfalls of power-law claims (Section 3.2), established the profound aggregation consequences of heavy tails via the single-big-jump principle (Section 3.3) and the generalized CLT (Section 3.4), and extended to multivariate/spatial settings via tail dependence (Section 3.5). Chapter 4 connected this entire apparatus to formal risk measurement: VaR's quantile-based simplicity but axiomatic failure (Section 4.1), CVaR's coherent remedy and its connection to the mean excess function (Section 4.2), the full axiomatic and representation theory of coherent/convex risk measures (Section 4.3) — revealing risk measurement and robust decision-making as two views of the same mathematics — spectral generalizations (Section 4.4), and the elicitability subtleties governing practical validation (Section 4.5).

Forward connections.

- *Volume II* (*Reliability Theory & Failure Processes*) applies the survival-function and hazard-rate concepts introduced for the GPD (Section 2.1) to lifetime data, develops the Wöhler/fatigue models whose stochastic versions connect to the first-passage and Lévy-stable distributions of Section 3.4, and revisits the GEV/Gumbel correspondence in the specific context of the (reliability) Weibull distribution (the notebox of Section 1.2).
- *Volume III* (*Robust Optimization & Decision under Uncertainty*) develops the worst-case-expectation representation of Section 4.3 into a full optimization framework (robust and distributionally robust optimization), and the decision-theoretic underpinnings previewed in Section 4.4's discussion of dual utility theory.
- *Volume IV* (*Cascading Failures & Network Resilience*) extends the multivariate tail dependence of Section 3.5 to network and spatial settings (Appendix C.2's max-stable processes), and the single-big-jump/diversification analysis of Section 3.3 and Appendix D.2 to systemic risk in financial and infrastructure networks.
- *Volume V* (*Rare Events & Large Deviations*) provides the large-deviation-theoretic foundations (Cramér's theorem, the Gärtner–Ellis theorem) that underlie both the entropic risk measure of Section 4.3 and the rare-event simulation techniques needed to estimate the tail probabilities and risk measures of this chapter when they are too small for direct Monte Carlo — a computational complement to the analytical extreme value methods developed here.
- *Volume VI* (*Antifragility, Resilience & Adaptive Capacity*) returns to the qualitative themes raised throughout this chapter — the asymmetry between typical and extreme behaviour, the limits of historical data for extrapolation, the deep uncertainty surrounding ξ — and develops formal frameworks for systems that are not merely robust to (i.e.

unaffected by) the extremes characterized here, but that can adapt to, or even benefit from, the volatility and disorder that heavy-tailed, extreme-value-governed processes inevitably produce.

A closing methodological reflection. Perhaps the single most important practical lesson of this volume's first chapter is the one most easily lost amid the mathematical formalism: *the uncertainty surrounding the shape parameter ξ — and hence surrounding every derived quantity (return levels, VaR, CVaR, moment existence) — is typically far larger than practitioners assume*, and this uncertainty does not diminish quickly with more data (Section 1.5's $m = 50$ example, Appendix A.1's wide confidence intervals, Appendix D.1's systematic understatement by plug-in estimates). Extreme value theory provides rigorous, principled tools for extrapolation beyond observed data — tools with no real alternative when the questions of interest (100-year floods, 1-in-10000 financial losses, ultimate tensile failure probabilities) concern events rarer than any feasible direct observation. But the rigour of the mathematics should not be mistaken for precision in the resulting numbers: *communicating the width of the uncertainty, not merely a point estimate, is an integral and non-optional part of any extreme value analysis*, and this principle — more than any specific formula — is the thread connecting this chapter to the risk, reliability, and resilience mathematics developed throughout the remainder of this series.

Appendix E

Extended Case Studies

Five integrated case studies applying the full machinery of Chapter 1 (Volume I) to realistic problems in hydrology, insurance, finance, engineering, and seismology.

E.1 Case Study: Design Flood Estimation for a River Levee

Problem statement

A river authority must determine the design height of a levee intended to provide protection against floods up to the "1-in-200-year" level, with an explicit requirement to quantify the uncertainty in this design level. Available data: 58 years of daily discharge records (in m^3/s) at the relevant gauging station.

Step 1: Exploratory analysis and stationarity check. Before any extreme value modelling, the analyst must assess whether the 58-year record can be treated as (approximately) stationary — a non-trivial question given documented upstream land-use changes (urbanization increasing impervious surface, hence runoff) over the period. A simple diagnostic: plot annual maximum discharge against year, and fit a linear trend. Suppose this trend is small and not statistically significant ($p = 0.21$ for the slope coefficient) — the analyst proceeds with a stationary model but flags the trend as a source of *model uncertainty* not captured by the within-model parameter uncertainty (Appendix D.1), to be revisited in sensitivity analysis.

Step 2: Method choice — block maxima vs. POT. With 58 years of daily data ($N \approx 21170$ observations), POT (Chapter 2) can in principle use far more data than the 58 annual maxima. However, daily river discharge is strongly autocorrelated (a flood event persists for days), so the extremal index θ (Appendix B.1) must be estimated and declustering applied. Given the practical complexity, suppose the analyst pursues *both* methods as a cross-check (Section 2.2's PBdH-guaranteed consistency).

Step 2a: Block maxima (GEV). The 58 annual maxima have sample mean $\bar{M} = 412\,\mathrm{m}^3/\mathrm{s}$, sample SD $s_M = 156$, sample skewness $\hat{\gamma}_1 = 0.71$. ML fitting (Section 1.5) yields:

$$\hat{\xi}_{\mathrm{GEV}} = 0.18\ (\mathrm{SE} = 0.12), \qquad \hat{\mu} = 345\ (\mathrm{SE} = 22), \qquad \hat{\sigma} = 132\ (\mathrm{SE} = 16).$$

Step 2b: POT (GPD). Mean excess plot analysis (Section 2.3) suggests threshold $u = 180\,\mathrm{m}^3/\mathrm{s}$, yielding $k = 215$ raw exceedances. Ferro–Segers extremal index estimate (Appendix B.1): $\hat{\theta} = 0.31$, suggesting an effective $k_{\mathrm{eff}} \approx 215 \times 0.31 \approx 67$ independent flood episodes — comparable to (though somewhat larger than) the 58 annual block maxima, consistent with most years having $\gtrsim 1$ flood episode above u. GPD ML fitting on declustered exceedances:

$$\hat{\xi}_{\mathrm{GPD}} = 0.21\ (\mathrm{SE} = 0.10), \qquad \hat{\beta} = 68\ (\mathrm{SE} = 11).$$

The two $\hat{\xi}$ estimates (0.18 from GEV, 0.21 from GPD) agree within their (substantial, SE $\approx$ 0.10–0.12) standard errors — consistent with PBdH (Section 2.2). A pooled estimate (e.g. inverse-variance weighted) might be $\hat{\xi} \approx 0.195$.

Step 3: 200-year design level. Using the POT parametrization (Section 2.5), $\lambda_{\mathrm{eff}} = k_{\mathrm{eff}}/58 \approx 67/58 \approx 1.16$ episodes/year. For $T = 200$: $\lambda T \approx 232$.

$$(\lambda T)^{\hat{\xi}} = 232^{0.195} = \exp(0.195 \times \ln 232) = \exp(0.195 \times 5.446) = \exp(1.062) \approx 2.892.$$

$$x_{200} = u + \frac{\hat{\beta}}{\hat{\xi}}[(\lambda T)^{\hat{\xi}} - 1] = 180 + \frac{68}{0.195} \times (2.892 - 1) = 180 + 348.7 \times 1.892 \approx 180 + 659.9 \approx 839.9\,\text{m}^3/\text{s}.$$

Step 4: Uncertainty quantification. Applying the delta method (Appendix A.1) with $\text{Var}(\hat{\xi}) \approx (0.10)^2 = 0.01$ (using the GPD SE), $\text{Var}(\hat{\beta}) \approx (11)^2 = 121$, and an estimated covariance $\text{Cov}(\hat{\xi}, \hat{\beta}) \approx -0.5$ (typical negative sign, magnitude estimated from the likelihood Hessian), the delta-method standard error on x_{200} might be computed (following the same procedure as Section 2.5's worked example, with the larger $\partial g / \partial \xi$ term dominating) as $\text{SE}(\hat{x}_{200}) \approx 95$.

Step 5: Recommendation and reporting. A 95% CI for the 200-year design level is approximately $840 \pm 1.96 \times 95 \approx (654, 1026)\,\text{m}^3/\text{s}$. The river authority faces a genuine decision under deep uncertainty (Chapter 11, Volume III foreshadowed): designing to the point estimate (840) accepts that there is roughly a 50% chance the "true" 200-year level exceeds this; designing to the upper confidence bound (≈ 1026) is far more conservative (and costly) but provides protection against a wider range of plausible true values. Best practice — increasingly mandated by engineering codes — is to report the full interval (or even the full posterior predictive distribution, Appendix D.1) to decision-makers, rather than presenting a single "design flood" number as if it were known with certainty. The non-stationarity flagged in Step 1, while not statistically significant in this dataset, would — if real — bias all of these estimates *downward* (urbanization increases flood peaks over time, so historical data underrepresents future risk), an additional one-directional consideration the authority should weigh, perhaps via an explicit non-stationary GPD model (Section 2.5's closing remarks) as a sensitivity check even if not adopted as the primary model.

E.2 Case Study: Catastrophe Reinsurance Pricing

Problem statement

A reinsurer is asked to price an excess-of-loss treaty covering aggregate annual losses from a specific peril (e.g. windstorm) in a region, for losses between a retention $R = \$50\text{M}$ and a limit $R + L = \$250\text{M}$ (i.e. the reinsurer pays $\min(\max(S - R, 0), L)$ for aggregate annual loss S). Available data: 35 years of annual aggregate loss data (inflation-adjusted), with 12 years having zero or negligible losses, and the remaining 23 years ranging from \$2M to \$310M.

Step 1: Frequency-severity decomposition. Rather than modelling annual aggregate loss S directly, decompose: $S = \sum_{i=1}^{N} X_i$ where N is the number of loss-causing events in a year (modelled as Poisson, $\hat{\lambda}_N \approx 23/35 \times$ (avg. events per loss year) — suppose years with losses average 1.4 events, giving $\hat{\lambda}_N \approx 0.92$/year) and X_i are individual event severities, modelled via GPD above a threshold (Chapter 2) for the upper tail and an empirical/lognormal body for smaller events.

Step 2: Severity tail modelling. The largest 40 individual event losses (pooled across all years) are used for GPD fitting above threshold $u = \$5\text{M}$:

$$\hat{\xi} = 0.45 \ (\text{SE} = 0.16), \qquad \hat{\beta} = \$8.2\text{M} \ (\text{SE} = 2.1).$$

Note $\hat{\xi} = 0.45$ is large — approaching the boundary $\xi = 0.5$ beyond which the GPD variance is infinite (Section 2.1) — consistent with the well-documented extremely heavy-tailed nature of catastrophe losses (a handful of "mega-events" dominate aggregate losses, the single-big-jump principle of Section 3.3 operating directly).

Step 3: Aggregate loss distribution via simulation. Given the frequency-severity model, $P(S > x)$ for large x (the relevant range, $x \in [\$50\text{M}, \$250\text{M}]$, well above individual-event thresholds in most cases) can be computed by Monte Carlo: simulate $N \sim \text{Poisson}(0.92)$, then N i.i.d. severities from the fitted severity model, sum, repeat $100000\times$.

By the single-big-jump principle (Section 3.3) — since $\hat{\xi} = 0.45$ places severities firmly in the Fréchet domain, hence subexponential — for x large relative to typical severities, $P(S > x) \approx \mathbb{E}[N] \times P(X > x) = 0.92 \times \overline{H}_{0.45,8.2}(x-5)$ (treating the threshold-exceedance probability as approximately the unconditional severity tail, a simplification valid when $u = \$5\text{M}$ is small relative to $x \sim \$100\text{M}+$).

Step 4: Expected layer loss. The reinsurer's expected payout is

$$\mathbb{E}[\min(\max(S - R, 0), L)] = \int_R^{R+L} P(S > x)\, dx.$$

Using the single-big-jump approximation, $P(S > x) \approx 0.92 \times (1 + 0.45(x-5)/8.2)^{-1/0.45} = 0.92 \times (1 + 0.0549(x-5))^{-2.222}$.

Numerical evaluation. At $x = R = 50$: $1 + 0.0549 \times 45 = 1 + 2.47 = 3.47$. $3.47^{-2.222} = \exp(-2.222 \times \ln 3.47) = \exp(-2.222 \times 1.244) = \exp(-2.764) \approx 0.0630$. So $P(S > 50) \approx 0.92 \times 0.0630 \approx 0.0580$ — i.e. roughly a 5.8% annual probability that aggregate losses exceed the $\$50\text{M}$ retention.

At $x = R + L = 250$: $1 + 0.0549 \times 245 = 1 + 13.45 = 14.45$. $14.45^{-2.222} = \exp(-2.222 \times \ln 14.45) = \exp(-2.222 \times 2.671) = \exp(-5.937) \approx 0.00264$. $P(S > 250) \approx 0.92 \times 0.00264 \approx 0.00243$ — about 0.24% annual probability of losses reaching the top of the layer.

The integral $\int_{50}^{250} P(S > x)dx$ (the expected layer loss) requires numerical integration of $0.92 \times (1 + 0.0549(x-5))^{-2.222}$ from 50 to 250. Substituting $w = 1 + 0.0549(x-5)$, $dw = 0.0549\, dx$, limits $w(50) = 3.47$ to $w(250) = 14.45$:

$$\int_{50}^{250} P(S > x)dx = \frac{0.92}{0.0549} \int_{3.47}^{14.45} w^{-2.222} dw = \frac{0.92}{0.0549} \times \frac{w^{-1.222}}{-1.222}\Big|_{3.47}^{14.45}.$$

$3.47^{-1.222} = \exp(-1.222 \times 1.244) = \exp(-1.520) \approx 0.2187$. $14.45^{-1.222} = \exp(-1.222 \times 2.671) = \exp(-3.264) \approx 0.0382$.

$$= \frac{0.92}{0.0549} \times \frac{0.0382 - 0.2187}{-1.222} = 16.76 \times \frac{-0.1805}{-1.222} = 16.76 \times 0.1477 \approx 2.475.$$

Expected layer loss $\approx \$2.475$M per year.

Step 5: Pricing and parameter-uncertainty loading. A pure risk premium would be $\approx \$2.475$M/year, but Appendix D.1's lesson applies forcefully here: $\hat{\xi} = 0.45$ with SE $= 0.16$ means the posterior/sampling distribution for ξ extends substantially into the region $\xi > 0.5$ (infinite-variance GPD) and even toward $\xi \approx 0.6$–0.7 within roughly one SE — in this region, the layer-loss integral above can increase dramatically (the integrand's exponent $-1/\xi$ approaches -1 from below, making the integral diverge as $\xi \to 1$, far beyond but in the direction of the estimated uncertainty). A responsible pricing approach must incorporate this parameter uncertainty — e.g. via the Bayesian predictive approach of Appendix D.1, or via a loading factor calibrated to the upper end of the $\hat{\xi}$ confidence interval — potentially increasing the technical price by 30–80% above the naive plug-in estimate of $\$2.475$M, before any commercial loading (expenses, profit margin, capital cost) is applied. This case study illustrates concretely why catastrophe reinsurance pricing is widely regarded as one of the

most demanding applications of extreme value theory: the combination of a very heavy tail ($\hat{\xi}$ near the variance-existence boundary), substantial parameter uncertainty (driven by the inherent rarity of the largest events), and a payoff structure (the reinsurance layer) that is itself most sensitive to exactly the tail region where uncertainty is greatest.

E.3 Case Study: Market Risk Capital Under Basel FRTB

Problem statement

A bank's trading desk holds a portfolio whose daily P&L (profit and loss) has been recorded for 1500 trading days (approximately 6 years). The bank must compute (a) $\text{VaR}_{0.99}$ and $\text{CVaR}_{0.975}$ (the levels specified under different historical Basel frameworks — Basel 2.5 used $\text{VaR}_{0.99}$, FRTB shifted to $\text{CVaR}_{0.975}$, chosen so that the two measures give comparable capital levels for Gaussian P&L, per Section 4.2's relationship), (b) backtest the VaR forecasts using the traffic-light approach (Section 4.5), and (c) assess whether the portfolio's loss distribution exhibits tail behaviour that would make EVT-based extrapolation advisable for stress scenarios beyond the directly observable quantiles.

Step 1: Empirical quantile estimation. With 1500 daily losses (negated P&L, so losses are positive), the empirical 99% quantile corresponds to the 15th largest loss (since $1500 \times 0.01 = 15$), and the 97.5% quantile to the 37.5th largest (interpolated between the 37th and 38th). Suppose these are $\widehat{\text{VaR}}^{\text{emp}}_{0.99} = \$4.2\text{M}$ and $\widehat{\text{VaR}}^{\text{emp}}_{0.975} = \$3.1\text{M}$.

Step 2: Is EVT needed? For $\text{VaR}_{0.99}$ with 1500 observations, 15 exceedances is a reasonable (if modest) sample for direct empirical quantile estimation — EVT extrapolation is not strictly necessary at *this* quantile. However, regulatory *stress testing* requirements often demand assessment at $\text{VaR}_{0.999}$ or beyond (corresponding to 1-2 exceedances in 1500 days, or fewer than 1 expected exceedance — direct empirical estimation is essentially impossible) — this is where Chapters 1–2's machinery becomes indispensable.

Step 3: GPD fit to the tail of daily losses. Mean excess plot analysis (Section 2.3) on the daily loss data suggests a threshold around the empirical 95% level, $u \approx \$2.0\text{M}$, yielding $k = 75$ exceedances ($1500 \times 0.05 = 75$). GPD ML fit:

$$\hat{\xi} = 0.28 \ (\text{SE} = 0.13), \qquad \hat{\beta} = \$0.85\text{M} \ (\text{SE} = 0.16).$$

$\hat{\xi} = 0.28$ indicates a moderately heavy tail — not unusual for daily financial P&L, reflecting well-documented "fat tails" relative to the Gaussian (which would correspond to $\xi = 0$, Gumbel domain, via the von Mises condition discussion of Section 1.4 — though as that section noted, even *some* Gumbel-domain distributions like log-normal can "look" heavy-tailed; a positive $\hat{\xi}$ here is direct evidence of genuine Fréchet-domain behaviour).

Step 4: EVT-based $\text{VaR}_{0.999}$. Using the POT return-level formula (Section 2.5) with $\lambda = 75/6 \approx 12.5$ exceedances/year (since 1500 days ≈ 6 years, and exceedances of the 95% daily threshold occur $\approx 12.5\times$/year by construction — actually 75 exceedances over 6 years $= 12.5$/year, consistent): for $\text{VaR}_{0.999}$ (annual, i.e. exceeded with probability 0.001 per year, $T = 1000$ years in the return-period sense, *or* equivalently we can work on a per-observation basis):

On a per-observation basis, $\text{VaR}_{0.999}$ corresponds to the level exceeded with probability 0.001 *per day*. Using $\overline{H}_{\hat{\xi},\hat{\beta}}(x-u) = \frac{0.001}{75/1500} = \frac{0.001}{0.05} = 0.02$:

$$(1 + 0.28(x-2)/0.85)^{-1/0.28} = 0.02 \ \Rightarrow \ 1 + 0.329(x-2) = 0.02^{-0.28}.$$

$$0.02^{-0.28} = \exp(-0.28 \times \ln 0.02) = \exp(-0.28 \times (-3.912)) = \exp(1.095) \approx 2.991.$$

$$0.329(x-2) = 1.991 \;\Rightarrow\; x - 2 \approx 6.052 \;\Rightarrow\; x \approx \$8.05\text{M}.$$

$\text{VaR}_{0.999} \approx \$8.05\mathbf{M}$ — roughly $1.9\times$ the empirical $\text{VaR}_{0.99}$ of $\$4.2$M, an extrapolation entirely reliant on the GPD tail model, since (as noted) only 1-2 exceedances of this level would be expected in the 6-year sample.

Step 5: CVaR at the regulatory level. $\text{CVaR}_{0.975}$ (Section 4.2): using the mean-excess formula with the GPD parameters at threshold $u = 2$:

$$e(\text{VaR}_{0.975}) = \frac{\hat{\beta} + \hat{\xi}(\text{VaR}_{0.975} - u)}{1 - \hat{\xi}} = \frac{0.85 + 0.28 \times (3.1 - 2)}{0.72} = \frac{0.85 + 0.308}{0.72} = \frac{1.158}{0.72} \approx 1.608.$$

$$\text{CVaR}_{0.975} \approx 3.1 + 1.608 \approx \$4.71\text{M}.$$

Note $\text{CVaR}_{0.975} \approx \$4.71\text{M} > \text{VaR}_{0.99}^{\text{emp}} = \$4.2\text{M}$ — under FRTB, the capital requirement based on $\text{CVaR}_{0.975}$ would *exceed* the old Basel 2.5 requirement based on $\text{VaR}_{0.99}$ for this portfolio, despite the lower confidence level (0.975 vs 0.99) — a consequence of CVaR's incorporation of tail magnitude (Section 4.2), which for this moderately heavy-tailed ($\hat{\xi} = 0.28$) portfolio more than compensates for the lower quantile level. This is a frequently-cited rationale for the regulatory shift: $\text{CVaR}_{0.975}$ is, for heavy-tailed portfolios, often *more* conservative than $\text{VaR}_{0.99}$, even though a naive (Gaussian-based) calibration intended them to be roughly equivalent.

Step 6: Backtesting (Section 4.5). Over the 1500-day sample, suppose the bank's *historical* $\text{VaR}_{0.99}$ model (which may differ from the in-sample empirical quantile used above, e.g. if it's a rolling-window or GARCH-based model) produced 19 violations. Under H_0 (correct calibration), expected violations $= 1500 \times 0.01 = 15$. Using a normal approximation to the binomial, $\text{SD} = \sqrt{1500 \times 0.01 \times 0.99} \approx 3.85$. The observed 19 is $(19 - 15)/3.85 \approx 1.04$ standard deviations above expectation — not statistically significant ($p \approx 0.30$, two-sided), placing the bank in the "green zone" (Section 4.5's traffic-light table, scaled appropriately for $N = 1500$ rather than 250) despite the realized violation rate ($19/1500 \approx 1.27\%$) being noticeably above the nominal 1%. This illustrates the generally *low power* of VaR backtests over realistic sample sizes — a model that is moderately miscalibrated (true coverage $\approx 98.7\%$ rather than 99%) may not be statistically distinguishable from a correctly-calibrated model for years, a limitation that motivates the joint VaR-CVaR backtesting approaches of Section 4.5's closing discussion, which can have greater power by also using information about the *magnitude* of violations, not merely their count.

E.4 Case Study: Fatigue Life Extreme Quantiles for a Critical Component

Problem statement

A manufacturer of a safety-critical mechanical component (e.g. a turbine blade) has fatigue-test data from $n = 45$ specimens tested to failure under constant-amplitude loading, with cycles-to-failure ranging from 1.2×10^5 to 8.7×10^6. The manufacturer must certify a "B10 life" (the cycle count at which 10% of the population is expected to have failed) with appropriate confidence bounds, and additionally assess the probability of failure within the first 10^5 cycles (the "infant mortality" / early-failure regime).

Connection to extreme value theory: minima, not maxima. Fatigue life is a *minimum*-type extreme: the component fails at the *weakest link* (the location of the most severe microstructural defect), and the B10 life corresponds to a *lower* quantile of the failure-cycle distribution — the analogue, for minima, of the upper quantiles (VaR, return levels) studied throughout this chapter for maxima. As noted in Section 1.1, minima are converted to maxima via $X \to -X$, and the three FTG domains (Section 1.2) apply equally, with the *Weibull* domain (Section 1.2, $\xi < 0$, bounded support from below by 0 for fatigue life — a component cannot fail in negative cycles) being the natural and, as Section 1.2's notebox discussed, the historically dominant choice for fatigue applications — though that notebox's distinction (extreme-value Weibull for *minima of the normalized data* vs. the reliability-Weibull as a *direct model* for the data itself) is precisely operative here.

Step 1: Direct (3-parameter) Weibull fit. Rather than fitting a GEV/GPD to *normalized* minima (which would require many repeated samples of 45 specimens, not available), the standard reliability-engineering approach fits a 3-parameter Weibull *directly* to the 45 cycles-to-failure $N_1, \dots, N_{45}$:

$$F(N) = 1 - \exp\left[-\left(\frac{N-\gamma}{\eta}\right)^m\right], \qquad N \geq \gamma,$$

with shape m, scale η, and location (threshold) γ (often $\gamma = 0$ is assumed for simplicity, giving the 2-parameter Weibull; we retain γ here as it materially affects the B10 life and early-failure probability, the quantities of interest).

Step 2: Parameter estimation. Suppose ML (or median-rank regression, a common alternative in reliability engineering, analogous in spirit to the PWM method of Section 2.4) yields:

$$\hat{m} = 2.3 \ (\text{SE} = 0.28), \qquad \hat{\eta} = 2.1 \times 10^6 \ (\text{SE} = 0.25 \times 10^6), \qquad \hat{\gamma} = 8 \times 10^4 \ (\text{SE} = 3 \times 10^4).$$

The non-zero $\hat{\gamma} = 8 \times 10^4$ is itself significant: it implies a "guaranteed minimum life" of 80000 cycles below which failure (according to this model) cannot occur — a strong and consequential modelling assumption that should be checked against the data (does the smallest observed failure, 1.2×10^5, sit comfortably above $\hat{\gamma} = 8 \times 10^4$? Yes, by a margin of 4×10^4, which is plausible but not overwhelming evidence given $\text{SE}(\hat{\gamma}) = 3 \times 10^4$).

Step 3: B10 life. $F(N_{B10}) = 0.10$:

$$N_{B10} = \gamma + \eta(-\log(1-0.10))^{1/m} = \gamma + \eta(-\log 0.9)^{1/m} = \gamma + \eta(0.10536)^{1/2.3}.$$

$$(0.10536)^{1/2.3} = \exp(\tfrac{1}{2.3} \times \ln 0.10536) = \exp(\tfrac{1}{2.3} \times (-2.250)) = \exp(-0.9783) \approx 0.3760.$$

$$N_{B10} = 8 \times 10^4 + 2.1 \times 10^6 \times 0.3760 \approx 8 \times 10^4 + 7.896 \times 10^5 \approx 8.70 \times 10^5.$$

$N_{B10} \approx 870000$ **cycles.**

Step 4: Uncertainty on B10 (delta method, Appendix A.1 applied to the Weibull). $N_{B10} = g(\gamma, \eta, m)$. Using the parameter SEs and an estimated covariance structure (3-parameter Weibull MLEs are typically substantially correlated, particularly $\hat{\gamma}$ and $\hat{\eta}$, with $\text{Corr}(\hat{\gamma}, \hat{\eta})$ often ≈ -0.6 to -0.8), suppose the delta-method SE on N_{B10} is $\approx 1.1 \times 10^5$. A 95% lower confidence bound (the quantity of regulatory interest — certification typically requires a *lower* bound, since understating B10 is conservative while overstating it is dangerous): $N_{B10}^{(0.95,\text{lower})} \approx 870000 - 1.645 \times 110000 \approx 870000 - 181000 \approx 689000$ (using a one-sided 95% bound, $z_{0.95} = 1.645$).

Step 5: Early-failure probability. $P(N < 10^5) = F(10^5)$:

$$F(10^5) = 1 - \exp\left[-\left(\frac{10^5 - 8 \times 10^4}{2.1 \times 10^6}\right)^{2.3}\right] = 1 - \exp\left[-\left(\frac{2 \times 10^4}{2.1 \times 10^6}\right)^{2.3}\right] = 1 - \exp\left[-(0.00952)^{2.3}\right].$$

$$(0.00952)^{2.3} = \exp(2.3 \times \ln 0.00952) = \exp(2.3 \times (-4.654)) = \exp(-10.70) \approx 2.25 \times 10^{-5}.$$

$$F(10^5) \approx 1 - \exp(-2.25 \times 10^{-5}) \approx 2.25 \times 10^{-5}.$$

The model predicts a probability of roughly 2.25×10^{-5} (about 1 in 44400) of failure before 10^5 cycles. This is a substantial extrapolation *below* the smallest observed failure (1.2×10^5) and is driven almost entirely by the fitted location parameter $\hat{\gamma} = 8 \times 10^4$ — a parameter estimated with SE $= 3 \times 10^4$, i.e. 37.5% relative uncertainty. The sensitivity of $F(10^5)$ to γ is extreme: if γ were instead, say, $\gamma = 5 \times 10^4$ (within roughly one SE of the estimate), then $(10^5 - 5 \times 10^4)/(2.1 \times 10^6) = 0.0238$, and $(0.0238)^{2.3} = \exp(2.3 \times (-3.738)) = \exp(-8.597) \approx 1.85 \times 10^{-4}$ — an **order-of-magnitude increase** in the predicted early-failure probability from a seemingly modest shift in γ well within its uncertainty band. **This is the fatigue-engineering analogue of the GEV/GPD shape-parameter sensitivity emphasized throughout this chapter**: a single parameter, estimated with substantial relative uncertainty from finite data, controls extrapolated tail (here, early-failure) probabilities to an extent that point estimates alone badly understate the true uncertainty — and for a safety-critical component, an order-of-magnitude uncertainty in the early-failure probability is not a mathematical curiosity but a central engineering concern, directly motivating the conservative (lower-confidence-bound) certification approach of Step 4, and — in current practice — increasingly motivating fully Bayesian treatments (Appendix D.1) of fatigue-life certification.

E.5 Case Study: Seismic Hazard Assessment — the Gutenberg–Richter Law as a Power Law

Problem statement

Seismic hazard for a region is to be assessed using 80 years of earthquake catalogue data (magnitudes $M \geq 3.0$, the catalogue's completeness threshold). The Gutenberg–Richter (GR) law states $\log_{10} N(\geq M) = a - bM$, where $N(\geq M)$ is the annual rate of earthquakes with magnitude at least M, and $b \approx 1$ is typical. The task: connect the GR law to the power-law/Pareto framework of Section 3.2, estimate the relevant parameters, assess goodness of fit (Section 3.2's recommended methodology), and compute the probability of a "maximum considered earthquake" (e.g. $M \geq 7.5$) within a 50-year design life.

Step 1: GR law as a Pareto distribution. The GR law $\log_{10} N(\geq M) = a - bM$ implies $N(\geq M) \propto 10^{-bM}$. Since seismic energy release E scales as $\log_{10} E \propto M$ (specifically $\log_{10} E = 1.5M + \text{const}$, the Richter-scale energy relation), $N(\geq M) \propto 10^{-bM} = E^{-b/1.5}$ — **the GR law is exactly a Pareto distribution for seismic energy**, with tail index $\alpha = b/1.5$. For $b = 1$, $\alpha = 2/3 < 1$ — recall from Section 3.1 that $\alpha < 1$ implies *infinite mean* energy release per earthquake (in the idealized, unbounded GR model) — a famous and physically significant fact: the "expected" energy released by an earthquake, under a pure GR/Pareto model with $b \approx 1$, is dominated entirely by the largest (rarest) events, to the point of formal divergence, directly illustrating the single-big-jump principle (Section 3.3) in a geophysical context.

Step 2: Estimating b (equivalently, α for magnitude itself). Working directly in magnitude units (rather than energy), $P(M > m \mid M > m_0) = 10^{-b(m-m_0)} = e^{-b \ln 10\,(m-m_0)}$ for $m \geq m_0 = 3.0$ — **magnitude, above the completeness threshold, is Exponentially distributed** with rate $\beta_{\text{exp}} = b \ln 10$. This is the Gumbel-domain / GPD-with-$\xi = 0$ case (Section 2.1) in magnitude units — a useful reminder that the *same* physical process (GR

law) is Pareto/Fréchet-domain in *energy* units but Exponential/Gumbel-domain in *magnitude* units, since magnitude is itself a logarithmic (hence "regularizing") transform of energy. The maximum-likelihood estimator for b (the "Aki estimator," directly analogous to the Hill estimator of Section 2.3, since both are MLEs for an exponential/Pareto tail) is

$$\hat{b} = \frac{\log_{10} e}{\bar{M} - m_0} = \frac{0.4343}{\bar{M} - 3.0},$$

where $\bar{M}$ is the mean magnitude of all events with $M \geq 3.0$ in the catalogue.

Example — Numerical estimation

Suppose the 80-year catalogue contains 3400 events with $M \geq 3.0$, with $\bar{M} = 3.42$. Then

$$\hat{b} = \frac{0.4343}{3.42 - 3.0} = \frac{0.4343}{0.42} \approx 1.034.$$

The annual rate of $M \geq 3.0$ events is $3400/80 = 42.5$/year, giving $\hat{a}$: from $\log_{10}(42.5) = a - \hat{b} \times 3.0 \Rightarrow a = \log_{10}(42.5) + 1.034 \times 3 = 1.629 + 3.102 = 4.731$.

Standard error of $\hat{b}$. As with the Hill estimator (Section 2.3), $\hat{b}$ is asymptotically normal with $\mathrm{SE}(\hat{b}) = \hat{b}/\sqrt{n}$ where $n = 3400$: $\mathrm{SE}(\hat{b}) \approx 1.034/\sqrt{3400} \approx 0.0177$. A remarkably precise estimate of b — reflecting the abundance of small-to-moderate earthquakes, in contrast to the sparse large-event data that plagued the GPD-tail estimates of earlier case studies. This is the seismological analogue of Section 2.3's observation that Hill-estimator precision improves with more order statistics (here, more catalogue events) — but, critically, $\hat{b}$ estimated this way characterizes the *bulk-to-moderate* magnitude range well-represented in the catalogue, and **extrapolating the same $\hat{b}$ to $M \geq 7.5$ (far beyond the largest events in most 80-year regional catalogues) assumes the GR law's single power-law (single-b) behaviour extends unbroken to the largest possible earthquakes** — an assumption increasingly questioned in seismology (the existence of a maximum possible magnitude $M_{\max}$, related to fault dimensions, would correspond to a *tapered* GR law, analogous to the power-law-with-exponential-cutoff discussion of Section 3.2).

Step 3: Probability of $M \geq 7.5$ in 50 years (pure GR extrapolation). Annual rate of $M \geq 7.5$: $\log_{10} N(\geq 7.5) = 4.731 - 1.034 \times 7.5 = 4.731 - 7.755 = -3.024$, so $N(\geq 7.5) = 10^{-3.024} \approx 9.46 \times 10^{-4}$/year — about once every 1057 years. Modelling occurrences as Poisson (Section 2.5's POT framework), $P(\text{at least one } M \geq 7.5 \text{ in 50 years}) = 1 - e^{-50 \times 9.46 \times 10^{-4}} = 1 - e^{-0.0473} \approx 0.0462$, roughly 4.6%.

Step 4: The $M_{\max}$ sensitivity. If, instead, the GR law is *tapered* with $M_{\max} = 8.0$ (a physically motivated cutoff based on regional fault geometry), the tapered Gutenberg–Richter (TGR) model modifies $N(\geq M)$ to decay faster than pure power-law as $M \to M_{\max}$ (analogous to the exponential-cutoff power law of Section 3.2). Even a cutoff at $M_{\max} = 8.0$ — comfortably above the $M \geq 7.5$ level of interest — can reduce $P(M \geq 7.5 \text{ in 50 yr})$ by a factor of 2–3× relative to the pure GR extrapolation, because the TGR's "tapering" begins to suppress probabilities meaningfully even somewhat below $M_{\max}$ itself. **The choice between pure GR and TGR (and the value of $M_{\max}$, itself highly uncertain and estimated from geological rather than statistical data) can change the 50-year probability of a major earthquake by a factor of** 2-3 — a model-uncertainty effect of the same order of magnitude as, but qualitatively distinct from, the *parameter*-uncertainty effects (in $\hat{\xi}$, $\hat{b}$, etc.) that have been the focus of most of this chapter's uncertainty discussions. This case study closes Appendix E by underscoring a theme first raised in Section 3.2 (the "fat middle" problem) and

recurring throughout: **statistical extrapolation of a power law fitted to abundant moderate data can be highly precise *and* substantially wrong, if the true tail behaviour departs from the fitted power law in ways that only become apparent — if at all — in the regime being extrapolated to.** No amount of data on $M \in [3, 6]$ earthquakes can, by itself, resolve whether $M_{\max} = 8.0$, 8.5, or does not exist at all; this is irreducible *model* uncertainty, distinct from (and not reducible by) the *parameter*-uncertainty quantification (confidence intervals on $\hat{b}$, $\hat{\xi}$, etc.) that has been this chapter's primary statistical focus, and addressing it requires — as in all five case studies of this appendix — combining the statistical machinery of Chapters 1–4 with domain-specific physical, engineering, or actuarial judgment that no purely data-driven procedure can supply.

Appendix F

Deferred Proofs and Comprehensive Formula Reference

Full derivations of results stated without complete proof in the main chapter, together with a consolidated reference of every key formula introduced in Chapter 1.

F.1 Full Proof: The Cauchy Functional Equation in the FTG Theorem

Overview

Section 1.2's proof sketch of the Fisher–Tippett–Gnedenko theorem invoked, without proof, the solution of the functional equations $\alpha(ts) = \alpha(t)\alpha(s)$ and $\beta(ts) = \alpha(t)\beta(s) + \beta(t)$. This section supplies the missing argument.

Lemma (Multiplicative Cauchy equation)

If $\alpha : (0,\infty) \to (0,\infty)$ satisfies $\alpha(ts) = \alpha(t)\alpha(s)$ for all $t, s > 0$ and α is monotonic (or merely measurable), then $\alpha(t) = t^{-\theta}$ for some $\theta \in \mathbb{R}$.

Proof. Define $\psi(x) = \log \alpha(e^x)$ for $x \in \mathbb{R}$. The functional equation becomes $\psi(x+y) = \log \alpha(e^{x+y}) = \log[\alpha(e^x)\alpha(e^y)] = \psi(x) + \psi(y)$ — the additive Cauchy equation. Under measurability (or monotonicity, or boundedness on any interval — any of the standard regularity conditions that rule out the pathological, non-measurable solutions constructible via the axiom of choice), the only solutions to $\psi(x+y) = \psi(x) + \psi(y)$ are $\psi(x) = cx$ for some constant c. [*Sub-proof*: $\psi(0) = \psi(0) + \psi(0) \Rightarrow \psi(0) = 0$. $\psi(nx) = n\psi(x)$ for $n \in \mathbb{N}$ by induction. $\psi(x/n) \times n = \psi(x)$, so $\psi(x/n) = \psi(x)/n$. Combining, $\psi(qx) = q\psi(x)$ for all $q \in \mathbb{Q}$. Continuity (from measurability + the Cauchy equation, by a standard argument — measurable solutions of Cauchy's equation are automatically continuous) extends this to $\psi(rx) = r\psi(x)$ for all $r \in \mathbb{R}$, hence $\psi(x) = \psi(1) \cdot x =: cx$.] Therefore $\alpha(e^x) = e^{cx}$, i.e. $\alpha(t) = t^c$ for $t > 0$. Writing $c = -\theta$ gives $\alpha(t) = t^{-\theta}$. ■

Lemma (The inhomogeneous equation for β)

Given $\alpha(t) = t^{-\theta}$, if $\beta : (0,\infty) \to \mathbb{R}$ satisfies $\beta(ts) = \alpha(t)\beta(s) + \beta(t)$ for all $t, s > 0$, then:

- If $\theta = 0$ (so $\alpha \equiv 1$): $\beta(ts) = \beta(s) + \beta(t)$ (additive Cauchy again), so $\beta(t) = c \log t$.
- If $\theta \neq 0$: $\beta(t) = \dfrac{c(1 - t^{-\theta})}{\theta}$ for some constant c.

Proof for $\theta \neq 0$. Define $\gamma(t) = \beta(t) + \frac{c}{\theta}$ for a constant c to be chosen; we seek c such that γ satisfies a multiplicative equation. The functional equation $\beta(ts) = t^{-\theta}\beta(s) + \beta(t)$, rewritten: $\beta(ts) - \frac{c}{\theta} = t^{-\theta}\left(\beta(s) - \frac{c}{\theta}\right) + \beta(t) + \frac{c}{\theta}(t^{-\theta} - 1) + \frac{c}{\theta}$, i.e. $\gamma(ts) = t^{-\theta}\gamma(s) + \beta(t) + \frac{c}{\theta}(t^{-\theta} - 1)$. For this to reduce to a clean multiplicative form $\gamma(ts) = t^{-\theta}\gamma(s) + \gamma(t) - \frac{c}{\theta} \cdot (\ldots)$, choose c such that $\beta(t) + \frac{c}{\theta}(t^{-\theta} - 1) = \gamma(t) \times (\text{const})$ — after some algebra (which we streamline by simply *verifying* the claimed solution rather than deriving it from scratch): substitute $\beta(t) = \frac{c(1-t^{-\theta})}{\theta}$ into the original equation and check it holds.

Verification. LHS: $\beta(ts) = \frac{c(1-(ts)^{-\theta})}{\theta} = \frac{c(1-t^{-\theta}s^{-\theta})}{\theta}$.

RHS: $\alpha(t)\beta(s) + \beta(t) = t^{-\theta} \cdot \frac{c(1-s^{-\theta})}{\theta} + \frac{c(1-t^{-\theta})}{\theta} = \frac{c}{\theta}\left[t^{-\theta}(1 - s^{-\theta}) + (1 - t^{-\theta})\right] = \frac{c}{\theta}\left[t^{-\theta} - t^{-\theta}s^{-\theta} + 1 - t^{-\theta}\right] = \frac{c}{\theta}\left[1 - t^{-\theta}s^{-\theta}\right]$.

LHS = RHS. ✓ ■

Completing the FTG proof. With $\alpha(t) = t^{-\theta}$ and $\beta(t) = \frac{c(1-t^{-\theta})}{\theta}$, the max-stability equation $G(x)^t = G(\alpha(t)x+\beta(t)) = G(t^{-\theta}x+\frac{c(1-t^{-\theta})}{\theta})$ must be solved for G. Taking $-\log(-\log G(x)) =: \Psi(x)$ (so $G(x) = \exp(-e^{-\Psi(x)})$, automatically giving a valid CDF shape provided Ψ is appropriately monotonic), the equation $G(x)^t = G(y)$ becomes $-\log(t \cdot e^{-e^{-\Psi(x)}}$'s exponent)... more directly: $G(x)^t = \exp(-te^{-\Psi(x)})$ and we need this to equal $\exp(-e^{-\Psi(y)})$ where $y = t^{-\theta}x + \frac{c(1-t^{-\theta})}{\theta}$, i.e. $te^{-\Psi(x)} = e^{-\Psi(y)}$, i.e. $\Psi(y) = \Psi(x) - \log t$.

For $\theta = 0$: $y = x + c\log t$ (taking the $\theta \to 0$ limit of $\beta(t)$, which is $c\log t$ by L'Hopital or direct substitution into the $\theta = 0$ case above), and $\Psi(x + c\log t) = \Psi(x) - \log t$ — setting $u = c\log t$ (so $\log t = u/c$), $\Psi(x + u) = \Psi(x) - u/c$, an affine Cauchy-type equation whose solution is $\Psi(x) = -x/c + \text{const}$, i.e. (absorbing constants into location/scale) $\Psi(x) = -x$, giving $G(x) = \exp(-e^{-x})$ — the **Gumbel** form, exactly as claimed in Section 1.2.

For $\theta \neq 0$: $y = t^{-\theta}x + \frac{c(1-t^{-\theta})}{\theta}$; writing $\Psi(x) = \theta\log(x - c/\theta)$ (a logarithmic ansatz, motivated by the scaling structure) and substituting confirms (after algebra analogous to the $\theta = 0$ case) that $\Psi(y) = \Psi(x) - \log t$ holds for this Ψ, up to the location shift c/θ that becomes the GEV's location parameter. This yields $G(x) \propto \exp(-(x - c/\theta)^\theta)$ type expressions which, after disentangling signs for $\theta > 0$ vs $\theta < 0$ and matching to the supports identified in Section 1.2, give the **Fréchet** ($\theta > 0$) and **Weibull** ($\theta < 0$) forms. ■ (We have presented the structure of the completing argument; a fully rigorous treatment additionally verifies the regularity/measurability conditions needed at each step, standard in the extreme value theory literature, e.g. de Haan and Ferreira's comprehensive treatment.)

F.2 Full Derivation: GEV Moments via the Gamma Function

Overview

Section 1.3 stated $\mathbb{E}[X^k]$ formulas in terms of $g_k = \Gamma(1 - k\xi)$ without derivation. We supply it here.

For the standard GEV ($\mu = 0, \sigma = 1$), $\xi \neq 0$, $\xi < 1/k$: $\mathbb{E}[X^k] = \int_{-1/\xi}^{\infty} x^k \, dG_\xi(x)$ (support adjusted appropriately for sign of ξ; we present the $\xi > 0$ case, support $x > -1/\xi$).

Substitute $z = (1 + \xi x)^{-1/\xi}$, so $x = \frac{z^{-\xi}-1}{\xi}$, and $G_\xi(x) = e^{-z}$, $dG_\xi = e^{-z}\,dz$ (note: as x ranges over $(-1/\xi, \infty)$, $z = (1+\xi x)^{-1/\xi}$ ranges over $(\infty, 0)$, i.e. $z \in (0, \infty)$ with reversed orientation, which the limits of integration must respect):

$$\mathbb{E}[X^k] = \int_0^\infty \left(\frac{z^{-\xi} - 1}{\xi}\right)^k e^{-z}\,dz.$$

For $k = 1$:

$$\mathbb{E}[X] = \frac{1}{\xi}\int_0^\infty (z^{-\xi} - 1)e^{-z}\,dz = \frac{1}{\xi}\left[\int_0^\infty z^{-\xi}e^{-z}dz - \int_0^\infty e^{-z}dz\right] = \frac{1}{\xi}\left[\Gamma(1-\xi) - 1\right],$$

using the definition of the Gamma function $\Gamma(s) = \int_0^\infty z^{s-1}e^{-z}dz$ with $s = 1 - \xi$ (requiring $1 - \xi > 0$, i.e. $\xi < 1$, for convergence — consistent with the $k = 1$ moment requiring $\xi < 1$). This gives $\mathbb{E}[X] = \frac{g_1 - 1}{\xi}$ with $g_1 = \Gamma(1 - \xi)$, as used in Section 1.3.

For $k = 2$: $\left(\frac{z^{-\xi}-1}{\xi}\right)^2 = \frac{z^{-2\xi}-2z^{-\xi}+1}{\xi^2}$, so

$$\mathbb{E}[X^2] = \frac{1}{\xi^2}\left[\Gamma(1-2\xi) - 2\Gamma(1-\xi) + \Gamma(1)\right] = \frac{1}{\xi^2}[g_2 - 2g_1 + 1],$$

requiring $\xi < 1/2$ for $g_2 = \Gamma(1-2\xi)$ to be finite. Then

$$\mathrm{Var}(X) = \mathbb{E}[X^2] - (\mathbb{E}[X])^2 = \frac{g_2 - 2g_1 + 1}{\xi^2} - \frac{(g_1-1)^2}{\xi^2} = \frac{g_2 - 2g_1 + 1 - (g_1^2 - 2g_1 + 1)}{\xi^2} = \frac{g_2 - g_1^2}{\xi^2},$$

confirming the formula used in Section 1.3.

Example — Higher moments and the general pattern

For general k, expanding $\left(\frac{z^{-\xi}-1}{\xi}\right)^k = \frac{1}{\xi^k}\sum_{j=0}^{k}\binom{k}{j}(-1)^{k-j}z^{-j\xi}$ and integrating term-by-term:

$$\mathbb{E}[X^k] = \frac{1}{\xi^k}\sum_{j=0}^{k}\binom{k}{j}(-1)^{k-j}\Gamma(1-j\xi),$$

requiring $\xi < 1/k$ (from the $j = k$ term, $\Gamma(1-k\xi)$, the most restrictive). For $k = 3$, the skewness involves $g_3 = \Gamma(1-3\xi)$, requiring $\xi < 1/3$ — exactly the boundary identified in Section 1.3's worked example ($\xi = 0.3 < 1/3 \approx 0.333$, third moment finite but only just).

F.3 Comprehensive Formula Reference

This section consolidates every major formula introduced in Chapter 1, organized by topic, for ease of reference.

Order statistics and maxima (Section 1.1).

$$F_{M_n}(x) = F(x)^n, \qquad f_{X_{(k)}}(x) = \frac{n!}{(k-1)!(n-k)!}F(x)^{k-1}[1-F(x)]^{n-k}f(x).$$

The three extreme value types (Section 1.2).

$$\Lambda(x) = \exp(-e^{-x}), \quad \Phi_\alpha(x) = \exp(-x^{-\alpha})\ (x>0), \quad \Psi_\alpha(x) = \exp(-(-x)^{\alpha})\ (x<0).$$

GEV (Section 1.3).

$$G_{\xi,\mu,\sigma}(x) = \exp\left\{-\left[1+\xi\frac{x-\mu}{\sigma}\right]^{-1/\xi}\right\}, \qquad x_p = \mu - \frac{\sigma}{\xi}\left[1-(-\log p)^{-\xi}\right].$$

$$\mathbb{E}[X] = \mu + \frac{\sigma}{\xi}(g_1 - 1), \quad \mathrm{Var}(X) = \frac{\sigma^2}{\xi^2}(g_2 - g_1^2), \quad g_k = \Gamma(1-k\xi),\ \xi < 1/k.$$

GPD (Section 2.1).

$$H_{\xi,\beta}(y) = 1 - \left(1+\xi\frac{y}{\beta}\right)^{-1/\xi}, \qquad e(v) = \frac{\beta+\xi v}{1-\xi}\ (\xi<1), \qquad h(y) = \frac{1}{\beta+\xi y}.$$

Threshold stability: excesses over $u_2 = u_1 + v$ are GPD$(\xi, \beta+\xi v)$.

PBdH and POT-GEV correspondence (Section 2.2, Appendix B.2).

$$\sigma_{\mathrm{GEV}} = \beta\lambda^{\xi}, \qquad \mu_{\mathrm{GEV}} = u + \frac{\sigma_{\mathrm{GEV}} - \beta}{\xi}.$$

Hill estimator (Section 2.3).

$$\hat{\xi}_{\mathrm{Hill}} = \frac{1}{k}\sum_{i=0}^{k-1}\left[\log X_{(N-i)} - \log X_{(N-k)}\right], \qquad \sqrt{k}(\hat{\xi}_k - \xi) \xrightarrow{d} \mathcal{N}(0,\xi^2)\ (\xi>0).$$

Return level (Section 2.5).

$$x_T = u + \frac{\beta}{\xi}\left[(\lambda T)^{\xi} - 1\right], \qquad (\xi = 0:\ x_T = u + \beta \log(\lambda T)).$$

Regular variation (Section 3.1).

$$f \in RV_\rho \iff f(x) = x^\rho L(x),\ L \in RV_0. \qquad \int_0^x f \sim \frac{x f(x)}{\rho + 1}\ (\rho > -1); \quad \int_x^\infty f \sim \frac{x f(x)}{|\rho + 1|}\ (\rho < -1).$$

Single big jump (Section 3.3). $F \in \mathcal{S} \Rightarrow P(S_n > x) \sim n P(X_1 > x)$.

Stable distributions / GCLT (Section 3.4). $\overline{F}(x) \sim c x^{-\alpha}$, $\alpha < 2 \Rightarrow \frac{S_n - b_n}{n^{1/\alpha}} \xrightarrow{d} X_\alpha$ (α-stable).

Tail dependence (Section 3.5).

$$\lambda_U = \lim_{q \to 1^-} P(Y > G^{-1}(q) \mid X > F^{-1}(q)). \qquad \text{Gaussian copula: } \lambda_U = 0\ \forall \rho < 1.$$

VaR and CVaR (Sections 4.1–4.2).

$$\mathrm{VaR}_q(L) = F_L^{-1}(q), \qquad \mathrm{CVaR}_q(L) = \frac{1}{1-q}\int_q^1 \mathrm{VaR}_u(L)\, du = \mathrm{VaR}_q(L) + e(\mathrm{VaR}_q(L)).$$

Coherence axioms (Section 4.3). Monotonicity, subadditivity, positive homogeneity, translation invariance. Representation: $\rho(L) = \sup_{Q \in \mathcal{Q}} \mathbb{E}_Q[L]$.

Spectral risk measures (Section 4.4).

$$\rho_\phi(L) = \int_0^1 \phi(u)\mathrm{VaR}_u(L)\, du, \qquad \phi \geq 0 \text{ non-decreasing}, \int_0^1 \phi = 1.$$

Elicitability (Section 4.5). VaR: pinball loss $S_q(x, y) = (q - \mathbf{1}\{y < x\})(y - x)$. CVaR: not elicitable alone; jointly elicitable with VaR (Fissler–Ziegel).

F.4 Glossary of Key Terms

Term	Brief definition / first reference
Block maxima	Maximum of each block of observations; basis of GEV fitting (§1.5)
CVaR / Expected Shortfall	Mean loss beyond VaR; coherent (§4.2)
Domain of attraction	Set of F converging to a given GEV type (§1.4)
Elicitability	Existence of a scoring function recovering a functional (§4.5)
Extremal index θ	Measure of clustering in dependent extremes (App. B.1)
Fréchet domain	$\xi > 0$; heavy/power-law tails (§1.2)
GEV	Generalized Extreme Value distribution (§1.3)
GPD	Generalized Pareto Distribution (§2.1)
Gumbel domain	$\xi = 0$; exponential-type tails (§1.2)
Hill estimator	MLE-type estimator of ξ in Fréchet domain (§2.3)
Mean excess function	$e(u) = \mathbb{E}[X - u \mid X > u]$ (§2.1)
Peaks-over-threshold (POT)	Modelling exceedances of a threshold via GPD (Ch. 2)
PBdH theorem	GPD as limit of threshold excesses (§2.2)
Power law	$\overline{F}(x) \sim Cx^{-\alpha}$ (§3.2)
Regular variation	$f(x) = x^{\rho}L(x)$, L slowly varying (§3.1)
Return level	Level with annual exceedance probability $1/T$ (§2.5)
Single big jump	$P(S_n > x) \sim nP(X_1 > x)$ for subexponential F (§3.3)
Spectral risk measure	Weighted integral of VaRs across confidence levels (§4.4)
Stable distribution	Limit of normalized sums for $\alpha \leq 2$ (§3.4)
Subexponential	$P(X_1 + X_2 > x)/P(X_1 > x) \to 2$ (§3.3)
Tail dependence λ_U	Conditional probability of joint extremes (§3.5)
Value-at-Risk (VaR)	q-quantile of loss distribution (§4.1)
Weibull domain	$\xi < 0$; bounded support (§1.2)

End of Volume I. This formula reference and glossary close Volume I of *Mathematics of Robustness, Risk & Failure*. The reader is encouraged to revisit Section D.4's synthesis for the conceptual threads connecting this volume's extreme-value-theoretic foundations to the reliability theory (Volume II), robust optimization (Volume III), network resilience (Volume IV), large deviations (Volume V), and resilience/antifragility (Volume VI) developed in the remainder of the series.

Appendix G

Extended Case Study: Operational Risk Capital Under the Loss Distribution Approach

A comprehensive, step-by-step application integrating every major method of Chapter 1 to a single realistic problem.

G.1 Problem Setup

Problem statement

A bank's operational risk department must estimate the 99.9% annual operational Value-at-Risk (the Basel-mandated confidence level for operational risk capital under the Advanced Measurement Approach / Loss Distribution Approach, historically) for a single business-line/event-type combination ("Internal Fraud, Retail Banking"). The internal loss database contains $n = 850$ individual loss events recorded over $T = 12$ years, with individual loss amounts ranging from the data-collection threshold of $\$10000$ to a maximum recorded loss of $\$47$ million. The department must: (1) model the loss severity distribution, with particular attention to the tail; (2) model the loss frequency; (3) combine these into an annual aggregate loss distribution; (4) compute $\text{VaR}_{0.999}$ and $\text{CVaR}_{0.999}$; and (5) quantify the uncertainty in the final capital number, integrating essentially every technique developed in this chapter.

G.2 Severity Modelling

Step 1: Exploratory analysis. The 850 losses (all $\geq \$10000$ by the collection threshold) have sample mean $\bar{X} = \$184000$, sample median $= \$52000$, and sample maximum $= \$47\text{M}$ — the enormous gap between mean and median, and between the maximum and the mean ($47\text{M}/0.184\text{M} \approx 255\times$), is the signature of extreme right-skew typical of operational loss data, and an immediate signal that EVT methods (rather than, say, a log-normal fit to the whole dataset) will be needed for the tail.

Step 2: Mean excess plot (Section 2.3). Computing $\hat{e}(u)$ for a grid of thresholds $u \in \{50000, 100000, 250000, 500000, 1000000, 2000000\}$:

u (\$)	$\hat{e}(u)$ (\$)	Exceedances k	Notes
50000	185000	412	near median; bulk of data
100000	298000	245	
250000	612000	98	
500000	1340000	41	
1000000	3100000	17	
2000000	7800000	6	very few exceedances

The ratio $\hat{e}(u)/u$ is roughly $3.7, 3.0, 2.4, 2.7, 3.1, 3.9$ across these thresholds — not perfectly stable, but without a clear monotonic trend that would indicate strong departure from GPD threshold-stability (Section 2.1) across this range. The slight increase at the highest thresholds ($u = 1\text{M}$, $u = 2\text{M}$) is based on very few exceedances (17, 6) and, per Section 2.3's discussion, should not be over-interpreted given the resulting standard errors (e.g. at $u = 2\text{M}$, $\text{SE}(\hat{e}) \approx 7.8\text{M}/\sqrt{6} \approx 3.18\text{M}$ — the point estimate 7.8M is barely more than 1 SE from the extrapolated linear trend from lower thresholds).

Step 3: Threshold selection. Balancing the considerations of Section 2.3 — the mean-excess

plot suggests reasonable linearity from $u = \$250000$ onward, while $u = \$500000$ retains $k = 41$ exceedances (a reasonable, if not large, sample for GPD fitting) and is high enough to be confidently in the "true tail" regime (well above the median of \$52000, by nearly an order of magnitude). The department selects $u = \$500000$.

Step 4: GPD fit. ML fitting (Section 2.4) to the $k = 41$ exceedances above \$500000 yields:

$$\hat{\xi} = 0.62 \text{ (SE} = 0.21), \qquad \hat{\beta} = \$680000 \text{ (SE} = 185000).$$

$\hat{\xi} = 0.62 > 0.5$: this is a critically important finding. Recall from Section 1.3/2.1 that $\xi \geq 0.5$ implies the GPD's *variance* is infinite, and $\xi \geq 1$ would imply infinite *mean*. With $\hat{\xi} = 0.62$, the point estimate places this severity distribution in the infinite-variance (but finite-mean, since $0.62 < 1$) regime — a result entirely consistent with, and indeed typical of, internal fraud loss data in retail banking, which is well-documented in the operational risk literature as among the heaviest-tailed of all Basel loss categories.

Step 5: Profile likelihood for ξ (Section 1.5's closing discussion, applied to the GPD). Given the consequential nature of whether $\xi < 0.5$ (finite variance) or $\xi \geq 0.5$ (infinite variance), the department computes a profile likelihood confidence interval for ξ rather than relying on the normal approximation $\hat{\xi} \pm 1.96$ SE. Suppose this yields a 95% profile likelihood CI of $(0.31, 1.04)$ — **spanning both the $\xi = 0.5$ (infinite variance) and $\xi = 1$ (infinite mean) boundaries**. This is a sobering but realistic finding for operational risk data with $k = 41$ tail observations: the data cannot statistically rule out infinite-mean severity at the 95% level, even though the point estimate (0.62) implies finite mean.

G.3 Frequency Modelling

Step 1: Annual event counts. Across 12 years, the annual counts of losses $\geq \$500000$ (the GPD threshold) are: $2, 1, 4, 2, 3, 5, 2, 1, 3, 4, 2, 6$ (sum $= 35$; note this differs from the $k = 41$ used for GPD fitting if, e.g., some exceedances were later identified as belonging to a 13th partial year or reclassified — for this case study we proceed with 35 over 12 years for frequency, illustrating that severity and frequency datasets need not be perfectly reconciled in practice, itself a source of model uncertainty).

Step 2: Poisson vs. Negative Binomial. Sample mean of annual counts $= 35/12 \approx 2.917$; sample variance ≈ 2.81 (computed from the data: deviations $(2-2.917)^2 + (1-2.917)^2 + \ldots$, summed and divided by 11). The variance (2.81) is slightly *below* the mean (2.917) — mild underdispersion relative to Poisson (which would have variance = mean). A formal dispersion test (e.g. a score test for overdispersion) likely fails to reject Poisson at conventional significance given the small sample ($n = 12$ years). The department adopts $N \sim \text{Poisson}(\hat{\lambda} = 2.917)$.

G.4 Aggregate Loss Distribution via Monte Carlo

Step 1: Simulation algorithm. For each of $M = 1000000$ simulated years:

1. Draw $N \sim \text{Poisson}(2.917)$.
2. For each of the N events, draw a severity X_i: with probability $(850 - 41)/850 \approx 0.952$, draw from the empirical distribution of the 809 losses below \$500000 (the "body"); with probability $41/850 \approx 0.0482$, draw from GPD$(\hat{\xi} = 0.62, \hat{\beta} = 680000)$ shifted by $+\$500000$ (the "tail").
3. Sum: $S = \sum_{i=1}^{N} X_i$.

This is a *splicing* approach (empirical body + parametric GPD tail), standard in operational risk LDA modelling and a direct application of the threshold/POT framework (Chapter 2) within a frequency-severity simulation.

Step 2: Numerical results (illustrative). From the $M = 10^6$ simulated annual aggregate losses S:

$$\text{VaR}_{0.999}(S) \approx \$38.6\text{M}, \qquad \text{CVaR}_{0.999}(S) \approx \$71.2\text{M}.$$

The ratio $\text{CVaR}_{0.999}/\text{VaR}_{0.999} \approx 1.84$ — consistent with (indeed, somewhat larger than) the ratio observed in Section 4.2's running example (≈ 1.51 at the 99% level), reflecting the heavier tail ($\hat{\xi} = 0.62$ here vs. 0.34 there) and the higher confidence level (99.9% vs. 99%, pushing further into the tail where the gap between VaR and CVaR widens further for heavy-tailed distributions, per the mean-excess linear-growth property of Section 2.1).

Step 3: Single-big-jump verification (Section 3.3). With $\hat{\xi} = 0.62$ (Fréchet domain, hence subexponential), the single-big-jump principle predicts $P(S > x) \approx \mathbb{E}[N] \times P(X > x)$ for x large. Check: at $x = \$38.6\text{M}$ (the computed $\text{VaR}_{0.999}$), $P(X > 38.6\text{M})$ from the GPD tail: $\overline{H}_{0.62,0.68}((38.6 - 0.5)) = (1 + 0.62 \times 38.1/0.68)^{-1/0.62} = (1 + 34.74)^{-1.613} = (35.74)^{-1.613}$. Compute $35.74^{-1.613} = \exp(-1.613 \times \ln 35.74) = \exp(-1.613 \times 3.5765) = \exp(-5.770) \approx 0.00313$. Then $P(X > 38.6\text{M} \mid \text{event}) \approx 0.0482 \times 0.00313 \approx 1.51 \times 10^{-4}$ (multiplying by the probability 0.0482 that any given event is a "tail" event in the first place, since the GPD survival function $\overline{H}$ is conditional on exceeding $u = 0.5\text{M}$). Then $\mathbb{E}[N] \times P(X > 38.6\text{M}) \approx 2.917 \times 1.51 \times 10^{-4} \approx 4.40 \times 10^{-4}$.

Compare to the target: $P(S > \text{VaR}_{0.999}) = 0.001$ by definition. The single-big-jump approximation gives 4.40×10^{-4}, about 44% of the target 0.001 — the approximation *under*-predicts $P(S > x)$ at $x = \text{VaR}_{0.999}$ by roughly a factor of 2, equivalently it would predict a *higher* $\text{VaR}_{0.999}$ than the simulation's $\$38.6\text{M}$ if used directly to solve $\mathbb{E}[N]P(X > x) = 0.001$ for x. This discrepancy reflects that, at the 99.9% level (only a 1000-year return period in annual terms), the "single big jump" approximation — while directionally correct and useful for understanding *why* the tail behaves as it does — has not yet fully kicked in: contributions from *two* moderately large events in the same year (rather than one dominant event) are still non-negligible at this confidence level, and only become asymptotically negligible (per Section 3.3's $P(S_n > x) \sim nP(X_1 > x)$, an asymptotic statement) at confidence levels considerably more extreme than 99.9%. This is an important practical caveat to the single-big-jump heuristic: **it describes the asymptotic regime, and at finite (even quite extreme, like 1-in-1000-year) confidence levels, "second-order" contributions from multiple large events can still constitute a near-factor-of-2 effect.**

G.5 Uncertainty Quantification: Putting It All Together

This final section integrates the uncertainty-quantification techniques of Appendix A.1 (delta method), Appendix D.1 (Bayesian/predictive), and the GPD shape-parameter sensitivity emphasized throughout the chapter, to bound the overall uncertainty in the $\$38.6\text{M}$ point estimate of $\text{VaR}_{0.999}$.

Source 1: Parameter uncertainty in $\hat{\xi}, \hat{\beta}$ (GPD). Given the wide profile-likelihood CI for ξ, $(0.31, 1.04)$ (Step 5 of G.2), re-running the Monte Carlo of Section G.4 with $\xi = 0.31$ (lower bound) and $\xi = 1.04$ (upper bound, holding β fixed at a correspondingly adjusted value to match the mean excess at $u = 0.5\text{M}$ — since ξ and β are correlated, simply varying ξ alone while holding β fixed would not respect the data) gives, illustratively:

$$\xi = 0.31: \ \text{VaR}_{0.999} \approx \$19.8\text{M}; \qquad \xi = 1.04: \ \text{VaR}_{0.999} \approx \$112\text{M} \ (\text{approaching divergence}).$$

A nearly $6\times$ range in $\text{VaR}_{0.999}$, from $\$19.8$M to $\$112$M, attributable solely to the 95% confidence interval on a single parameter, ξ. At ξ values approaching and exceeding 1, the underlying severity distribution's mean itself becomes ill-defined/infinite in the limit, and Monte Carlo estimates of $\text{VaR}_{0.999}$ become highly unstable (large simulation error even with $M = 10^6$ replications, since the result is increasingly dominated by the single largest simulated severity draw — an extreme manifestation of the single-big-jump principle operating *within* the simulation itself).

Source 2: Threshold choice ($u = \$500000$ vs. alternatives). Re-running the entire analysis (Sections G.2–G.4) with $u = \$250000$ (yielding $k = 98$ exceedances, potentially more stable $\hat{\xi}$ but with the GPD approximation possibly less accurate, per Section 2.3's bias-variance discussion) might yield $\hat{\xi} \approx 0.51$ (notably lower than 0.62) and a correspondingly lower $\text{VaR}_{0.999} \approx \$26$M. The choice of threshold alone — a decision made by the analyst, not dictated uniquely by the data — shifts the capital estimate by nearly $\$13$M ($38.6$M vs. 26M), a 33% relative difference.

Source 3: Frequency model uncertainty. If, instead of Poisson, a Negative Binomial frequency model (allowing for overdispersion, perhaps justified by a longer historical record showing more variability than the 12-year sample suggests) with the same mean 2.917 but variance $2\times$ higher were used, $\text{VaR}_{0.999}(S)$ would increase — higher frequency variance increases the probability of years with multiple large losses, directly relevant given Section G.4 Step 3's finding that multi-event years are non-negligible at the 99.9% level. The magnitude of this effect (perhaps a further 10–20% increase) compounds with Sources 1 and 2.

Final synthesis. The point estimate $\text{VaR}_{0.999} \approx \$38.6$M is a single number that will, in practice, be reported to regulators and used to set binding capital requirements. This appendix's uncertainty analysis reveals that this number is surrounded by genuine uncertainty — spanning, by Source 1 alone, roughly $\$20$M to over $\$100$M at standard confidence levels, before even considering Sources 2 and 3 (threshold choice, frequency model) which are not even *statistical* uncertainties in the usual sense but *modelling choices* with no objectively "correct" answer. **This is not a deficiency of the analysis; it is an accurate reflection of the genuine, irreducible difficulty of estimating a 1-in-1000-year quantile from roughly a decade of data containing a few dozen relevant tail observations.** The mathematical machinery of this entire chapter — the GEV/GPD families (Sections 1.3, 2.1), the PBdH correspondence (Section 2.2), threshold selection (Section 2.3), estimation methods and their standard errors (Section 2.4, Appendix A.1), the single-big-jump principle and its finite-sample caveats (Section 3.3, Section G.4 Step 3), and the risk-measure framework connecting all of this to a regulatory capital number (Chapter 4) — does not *eliminate* this uncertainty. What it does is make the uncertainty *visible, quantifiable, and communicable* — transforming an unavoidably difficult inference problem from one where the difficulty is hidden behind a single confident-looking number, into one where the difficulty is explicit, its sources are identified, and its consequences for decision-making (here, regulatory capital adequacy) can be reasoned about openly. This, ultimately, is the central practical contribution of extreme value theory and coherent risk measurement to the broader subject of this volume — and of the series of which it is the first installment.

Appendix H

Historical Notes and Further Reading

The development of extreme value theory and coherent risk measurement, and pointers for deeper study.

H.1 Historical Development of Extreme Value Theory

The mathematical study of extremes has origins distinct from, and in some ways predating, mainstream probability theory's focus on sums and averages. Maurice Fréchet's 1927 paper identified the possible limiting distributions for normalized maxima, motivated by problems in actuarial science. Ronald Fisher and L.H.C. Tippett's 1928 paper, working independently and from a more applied (specifically, strength-of-materials testing) perspective, arrived at an overlapping classification — the three types now bearing their names alongside Boris Gnedenko, who in 1943 provided the first fully rigorous proof of the complete classification theorem under precise regularity conditions, in what we now call the Fisher–Tippett–Gnedenko theorem (Section 1.2).

The unification of the three types into the single-parameter GEV family (Section 1.3) is generally attributed to the 1955 work of Emil Gumbel's contemporaries, though Gumbel himself — whose name is given to the $\xi = 0$ case — focused primarily on that special case in his influential 1958 book *Statistics of Extremes*, which brought extreme value methods to a wide applied audience, particularly in hydrology and engineering. The peaks-over-threshold / GPD framework (Chapter 2) developed somewhat later: James Pickands' 1975 paper and the independent work of A.A. Balkema and Laurens de Haan (also 1974) established the GPD as the universal limit for threshold excesses (Section 2.2), completing the "second pillar" of the theory roughly five decades after the first.

The regular variation framework (Chapter 3, Section 3.1) draws on the foundational work of Jovan Karamata in the 1930s, originally developed in the context of Tauberian theorems (relating the behaviour of a function/series to that of its transform) rather than probability per se — its application to extreme value theory came somewhat later, with the comprehensive treatment by de Haan in the 1970s establishing regular variation as the natural language for domains of attraction (Section 1.4). The theory of stable distributions (Section 3.4) has an even longer pedigree, with Paul Lévy's foundational work in the 1920s-1930s on what are now called Lévy-stable distributions, though their application to financial returns — and the associated controversies over whether financial data genuinely exhibits infinite-variance behaviour — is most strongly associated with Benoît Mandelbrot's work beginning in the early 1960s, predating by decades the more recent "fat tails" literature in quantitative finance.

H.2 Historical Development of Coherent Risk Measures

The axiomatic coherent risk measure framework (Section 4.3) is comparatively recent: the foundational paper of Philippe Artzner, Freddy Delbaen, Jean-Marc Eber, and David Heath, circulated from 1997 and published in 1999, crystallized a set of axioms that — while individually unsurprising — had not previously been assembled into a unified framework against which existing practice (dominated by VaR, Section 4.1) could be critically evaluated. The immediate identification of VaR's subadditivity failure (the counterexample structure of Section 4.1 traces to this period) was, at the time, a genuinely controversial claim, given VaR's near-universal adoption following its promotion by J.P. Morgan's RiskMetrics methodology in the mid-1990s and its subsequent enshrinement in the Basel I and Basel II capital accords.

The CVaR/Expected Shortfall alternative (Section 4.2), while mathematically simple once stated, took considerably longer to achieve regulatory adoption — the shift from $\text{VaR}_{0.99}$ to $\text{CVaR}_{0.975}$ as the basis for market risk capital was formalized only in the Basel Committee's Fundamental Review of the Trading Book (FRTB), with consultation papers beginning around 2012 and final implementation considerably later, i.e. roughly *fifteen years* after the theoretical case against VaR's coherence was first made — an instructive illustration of the gap between mathematical/theoretical results and their translation into binding regulatory practice, a gap influenced by legitimate practical concerns (including, as Section 4.5 discussed, CVaR's non-elicitability, not resolved until Gneiting's 2011 paper and the subsequent Fissler–Ziegel joint-elicitability result of 2016 — meaning the *practical validation* question for CVaR was not even fully resolved until *after* FRTB's adoption process had substantially begun).

The Kusuoka representation (Section 4.3's closing notebox) and the broader theory of law-invariant coherent risk measures developed through the 2000s, with spectral risk measures (Section 4.4) and their connection to Yaari's 1987 dual utility theory providing a decision-theoretic grounding that — as with the GPD/GEV unification decades earlier in extreme value theory — arrived as a *unifying synthesis* some years after the individual pieces (CVaR, distortion functions, rank-dependent utility) had each been developed somewhat independently across the actuarial, financial, and decision-theory literatures.

H.3 Connections to Adjacent Fields Not Covered in This Volume

Several substantial bodies of work connect closely to this chapter's content but fall outside its scope, either because they are developed in later volumes of this series or because they represent genuinely separate (though related) research traditions:

Records and record statistics. The theory of *records* — the sequence of successive maxima $X_{(1)} = X_1, X_{(2)} = \max(X_1, X_2), \ldots$ and the times at which new records occur — is closely related to extreme value theory (indeed, the distribution of the k-th record value is related to the GEV) but constitutes its own substantial literature, with applications to sports statistics, climate "record-breaking" analysis, and reliability (the time to a "record" failure in a renewal process, connecting to Volume II's renewal theory, Chapter 7).

Extremes in time series: GARCH and stochastic volatility. This volume's treatment of dependent extremes (Appendix B.1's extremal index) is necessarily brief. A substantial literature addresses extremes specifically in the context of GARCH-type and stochastic volatility models widely used for financial returns, where the extremal index can be computed (semi-)analytically for specific model parametrizations, and where the interaction between heavy-tailed innovations (Chapter 3) and volatility clustering produces extremal dependence structures distinct from — though related to — the simple max-autoregressive example of Appendix B.1.

Extreme value theory for machine learning and high-dimensional data. An active and rapidly developing area applies extreme value concepts to anomaly detection, to the analysis of neural network training dynamics (e.g. characterizing the tail behaviour of gradient distributions), and to high-dimensional generalizations of the multivariate extremes of Section 3.5 (where the "angular measure" ν must be estimated in dimensions $d \gg 2$, a problem of substantial current statistical research interest). These applications, emerging largely after the classical theory presented in this chapter was established, represent some of the most active current research directions building on this volume's foundations.

Climate extremes and non-stationarity. Section 2.5's brief discussion of non-stationary ex-

tensions barely scratches the surface of a large and policy-relevant literature on detecting and modelling trends in climate extremes (temperature, precipitation, sea level) under anthropogenic climate change — a literature that combines this chapter's statistical methodology with climate model output, downscaling techniques, and the attribution science that seeks to quantify the extent to which observed extremes can be attributed to specific causes (a question of *causal* inference layered on top of the *statistical* extreme value modelling developed here).

H.4 Suggested Further Reading

For the reader wishing to deepen their study of the topics in this chapter, the following (described qualitatively, without specific bibliographic citation, consistent with this volume's emphasis on self-contained mathematical development) areas of the literature are recommended:

- **Foundational textbook treatments** of univariate extreme value theory typically develop the FTG and PBdH theorems (Sections 1.2, 2.2) with full rigor, including the regularity conditions glossed over in this chapter's proof sketches (Appendix F.1), and provide comprehensive treatments of the statistical estimation methods (Sections 1.5, 2.3-2.4) with extensive simulation studies comparing ML, PWM, and Bayesian approaches across a wide range of ξ values.
- **Regular variation and its applications** are developed comprehensively in dedicated monographs that treat the Tauberian-theorem origins of the theory (Section 3.1) alongside its probabilistic applications, providing the technical foundation for the uniform convergence theorem and other results used but not fully proven in this chapter.
- **Multivariate and spatial extremes** (Section 3.5, Appendix C.2) are the subject of an active research literature on max-stable process models, copula-based dependence modelling, and their estimation — particularly relevant for the network and systemic risk applications previewed in this volume and developed in Volume IV.
- **Quantitative risk management** textbooks integrate the extreme value and copula methodology of this chapter with the broader toolkit of financial risk modelling (including the GARCH/stochastic volatility connections of Section H.3), providing the bridge between the pure mathematics of this chapter and its implementation in banking and insurance practice.
- **Coherent risk measures and their axiomatic foundations** (Section 4.3) are developed in the mathematical finance literature with full treatment of the representation theorems, their connections to no-arbitrage pricing theory, and the relationship to robust optimization (previewed here, developed fully in Volume III).

A note on mathematical maturity and prerequisites. This chapter has assumed familiarity with measure-theoretic probability (convergence in distribution, characteristic functions), real analysis (regular variation, the Gamma function), and basic statistical inference (maximum likelihood, asymptotic normality, confidence intervals). Readers comfortable with this material should find the chapter's proofs and derivations (including the appendices) largely self-contained; readers for whom this background is less familiar may benefit from reviewing these prerequisites before proceeding to Volume II, which builds directly on the survival-function and hazard-rate concepts introduced here (Section 2.1) in the substantially more applied context

of reliability engineering — a context where, as Appendix E.4's fatigue case study illustrated, the same mathematical structures (Weibull distributions, threshold parameters, extreme quantile extrapolation with attendant uncertainty) reappear with a different vocabulary and a different — though closely related — set of practical concerns.

Appendix I

Non-Stationary Extremes and Time-Varying Risk Measures

A rigorous treatment of covariate-dependent GEV/GPD models, the multiple competing notions of "return period" under non-stationarity, and their consequences for dynamic risk measurement.

I.1 Covariate-Dependent GEV and GPD Models

Overview

Sections 1.5 and 2.5 assumed the GEV/GPD parameters (ξ, μ, σ) or (ξ, β) are constant over time. This appendix develops the general framework in which these parameters depend on covariates $\mathbf{z}_t$ (time, climate indices, economic regime indicators), derives the resulting likelihood, and — crucially — shows how the very concept of "the" extreme value distribution becomes time-indexed, with consequences that ripple through every downstream quantity (return levels, VaR, CVaR) developed in this volume.

Definition (Non-stationary GEV)

For block maxima M_t observed at times $t = 1, \dots, m$ with covariates $\mathbf{z}_t \in \mathbb{R}^p$, the non-stationary GEV model specifies

$$M_t \sim G_{\xi_t, \mu_t, \sigma_t}, \qquad \mu_t = \mathbf{x}_t^\top \boldsymbol{\beta}_\mu, \quad \log \sigma_t = \mathbf{x}_t^\top \boldsymbol{\beta}_\sigma, \quad \xi_t = \xi \text{ (often held constant)},$$

for design vectors $\mathbf{x}_t$ built from $\mathbf{z}_t$ (e.g. $\mathbf{x}_t = (1, t)$ for a linear trend in μ, or $\mathbf{x}_t = (1, \text{ENSO}_t)$ for dependence on a climate index).

Why ξ_t is (usually) held constant. The shape parameter ξ governs the qualitative *type* of tail behaviour (Gumbel/Fréchet/Weibull, Section 1.2) and the existence of moments (Section 1.3). Allowing ξ itself to vary with covariates is statistically extremely demanding — ξ is already, as repeatedly emphasized in this chapter, the parameter estimated with the largest relative uncertainty from a *stationary* sample; permitting it to vary introduces additional parameters whose identification requires correspondingly more data, often unrealistically so. The near-universal practical compromise — ξ constant, (μ, σ) covariate-dependent — reflects a judgment that location/scale non-stationarity (the "where" and "how spread out" of the distribution shifting) is both more plausible physically (e.g. a warming climate shifts the location of the temperature-maxima distribution) and more estimable statistically than shape non-stationarity (a fundamental change in tail *type*).

The non-stationary likelihood. The log-likelihood for $\boldsymbol{\theta} = (\boldsymbol{\beta}_\mu, \boldsymbol{\beta}_\sigma, \xi)$ is simply the sum, over $t = 1, \dots, m$, of the GEV log-density (Section 1.3) evaluated at (ξ, μ_t, σ_t) with μ_t, σ_t as functions of $\boldsymbol{\theta}$ and $\mathbf{x}_t$:

$$\ell(\boldsymbol{\theta}) = \sum_{t=1}^{m} \log g_{\xi, \mu_t, \sigma_t}(M_t).$$

This is maximized numerically exactly as in the stationary case (Section 1.5), with the additional complexity that the constraint $1 + \xi(M_t - \mu_t)/\sigma_t > 0$ must hold for *every* t — a potentially more restrictive set of constraints than the single-threshold case, particularly if σ_t varies substantially (e.g. if variance is increasing over time, the constraint is most binding for the *earliest* observations, where σ_t is smallest).

Example — Worked example: detecting a trend in annual maximum temperature

Consider $m = 60$ years of annual maximum daily temperature, with covariate $z_t = t - 30.5$ (centered year index, $t = 1, \ldots, 60$, so $z_t \in [-29.5, 29.5]$). Fit two models:
Model 0 (stationary): $\mu_t = \mu_0$, $\log \sigma_t = \log \sigma_0$, giving $\hat{\xi}_0 = -0.08$ (SE 0.09), $\hat{\mu}_0 = 34.2$°C (SE 0.31), $\hat{\sigma}_0 = 2.1$°C (SE 0.21). Log-likelihood $\ell_0 = -128.4$.
Model 1 (linear trend in μ): $\mu_t = \mu_0 + \mu_1 z_t$, $\log \sigma_t = \log \sigma_0$ (constant), giving $\hat{\xi}_1 = -0.06$ (SE 0.09), $\hat{\mu}_0 = 34.3$ (SE 0.30), $\hat{\mu}_1 = 0.028$°C/yr (SE 0.011), $\hat{\sigma}_0 = 2.0$ (SE 0.19). Log-likelihood $\ell_1 = -125.7$.
Likelihood ratio test. $2(\ell_1 - \ell_0) = 2(-125.7 - (-128.4)) = 2 \times 2.7 = 5.4$. Under H_0 (no trend, $\mu_1 = 0$), this statistic is asymptotically χ^2_1. The critical value at 5% significance is $\chi^2_{1,0.95} = 3.841$. Since $5.4 > 3.841$, **the trend is statistically significant at the 5% level** ($p \approx 0.020$).
Interpretation. The estimated trend $\hat{\mu}_1 \approx 0.028$°C/year corresponds to a total shift of $\hat{\mu}_1 \times 59 \approx 1.65$°C over the 60-year record (comparing the first and last years) — a substantial shift relative to $\hat{\sigma}_0 \approx 2.0$°C, of nearly 0.83 standard deviations. **This single example already illustrates why "return level" becomes ambiguous**: the GEV distribution governing annual maxima in year 1 (with $\mu_1 \approx 34.3 - 0.028 \times 29.5 \approx 33.5$°C) is materially different from that governing year 60 (with $\mu_{60} \approx 34.3 + 0.028 \times 29.5 \approx 35.1$°C) — a "100-year return level" computed from which distribution? This question is addressed in Section I.2.

I.2 Three Competing Notions of Return Period Under Non-Stationarity

Overview

Under stationarity, "the T-year return level x_T" has a single, unambiguous meaning (Section 2.5): the level with annual exceedance probability $1/T$, equivalently the level whose expected waiting time to first exceedance is T years. Under non-stationarity, these two characterizations — "probability $1/T$ in a given year" and "expected waiting time T" — generally correspond to *different numerical values*, and a third notion (the "effective return level," relevant for fixed-design-life infrastructure) provides yet a different answer. This appendix section defines all three precisely and quantifies the discrepancies.

Definition (Three return period concepts)

Let G_t denote the (time-varying) GEV CDF for the annual maximum in year t.
(1) Fixed-probability level, $x_T^{(t_0)}$: the level x such that $1 - G_{t_0}(x) = 1/T$, computed using the distribution *at a specific reference year* t_0 (e.g. the present, or the design year).
(2) Expected waiting time, x_T^{EWT}: the level x such that the *expected number of years until first exceedance*, starting from year 1 and with G_t evolving as specified, equals T:

$$\mathbb{E}\left[\min\{t \geq 1 : M_t > x\}\right] = T.$$

(3) Design-life exceedance probability, $p^{\mathrm{DL}}(x; t_0, L)$: for an asset with design life L years beginning in year t_0, the probability of at least one exceedance of x during $[t_0, t_0 + L)$:

$$p^{\mathrm{DL}}(x; t_0, L) = 1 - \prod_{t=t_0}^{t_0+L-1} G_t(x).$$

The "T-year design level for an L-year design life" is then the x such that $p^{\mathrm{DL}}(x; t_0, L) = L/T$ (a convention chosen so that, under stationarity, this reduces to the familiar x_T).

Why these differ. Under stationarity ($G_t \equiv G$ for all t), all three definitions coincide: $1 - G(x) = 1/T$ gives $x_T^{(t_0)}$ for any t_0 (no dependence on reference year); the expected waiting time for a Bernoulli$(1/T)$ sequence of yearly exceedance indicators is exactly T (geometric distribution mean); and $p^{\mathrm{DL}}(x; t_0, L) = 1 - (1 - 1/T)^L \approx L/T$ for $L \ll T$. Under non-stationarity, none of these equivalences holds in general.

Example — Numerical illustration: the three return levels under a linear trend

Continuing the temperature example: suppose $\hat{\xi} = -0.06$, $\hat{\sigma}_0 = 2.0$, and $\mu_t = 33.5 + 0.028(t-1)$ for $t = 1, 2, \ldots$ (re-parametrizing so $t = 1$ corresponds to the start of a new 100-year planning horizon, with $\mu_1 = 33.5$ matching the early-record estimate from the worked example above). We seek the "100-year" level under each definition.

(1) Fixed-probability, $t_0 = 1$: Solve $1 - G_{\xi=-0.06,\mu=33.5,\sigma=2.0}(x) = 0.01$. Using the GEV quantile formula (Section 1.3) with $p = 0.99$, $(-\log 0.99)^{-\xi} = (0.01005)^{-(-0.06)} = (0.01005)^{0.06}$. Compute $(0.01005)^{0.06} = \exp(0.06 \times \ln 0.01005) = \exp(0.06 \times (-4.600)) = \exp(-0.276) \approx 0.759$.

$$x_{100}^{(1)} = \mu - \frac{\sigma}{\xi}[1 - 0.759] = 33.5 - \frac{2.0}{-0.06} \times 0.241 = 33.5 + 33.33 \times 0.241 \approx 33.5 + 8.03 \approx 41.53°\mathrm{C}.$$

(2) Expected waiting time. This requires solving $\sum_{t=1}^{\infty} t \cdot \overline{G}_t(x) \prod_{s<t} G_s(x) = T = 100$ for x — no closed form, but can be computed numerically. Because μ_t is *increasing* (warming trend), $\overline{G}_t(x)$ is *increasing* in t for fixed x (a fixed temperature becomes progressively more likely to be exceeded as the climate warms) — exceedance becomes progressively more likely in later years, which tends to *shorten* the expected waiting time relative to the stationary ($t_0 = 1$) calculation. Numerically, this might give $x_{100}^{\mathrm{EWT}} \approx 40.95°\mathrm{C}$ — **about** $0.6°\mathrm{C}$ **lower** than $x_{100}^{(1)} = 41.53°\mathrm{C}$: a level that, if held fixed, would (on average) be first exceeded within 100 years is *lower* than the level that has a 1% chance of exceedance in year 1 alone, precisely because the exceedance probability grows in later years.

(3) Design-life, $L = 50$ years, $t_0 = 1$. Solve $p^{\mathrm{DL}}(x; 1, 50) = 50/100 = 0.5$, i.e. $\prod_{t=1}^{50} G_t(x) = 0.5$. Since μ_t increases over the 50-year window from 33.5 to $33.5 + 0.028 \times 49 \approx 34.87$, this is an "average" over a range of distributions. Numerically, this might give $x_{100,50}^{\mathrm{DL}} \approx 41.05°\mathrm{C}$ — again **below** $x_{100}^{(1)}$, for similar reasons, but by a smaller margin than the EWT calculation (since the design-life calculation only "looks ahead" 50 years, not indefinitely as the EWT calculation implicitly does).

Summary table:

Definition	Value (°C)	Relative to $x_{100}^{(1)}$
(1) Fixed-probability ($t_0 = 1$)	41.53	— (reference)
(2) Expected waiting time	40.95	$-0.58°\mathrm{C}$
(3) Design-life ($L = 50$)	41.05	$-0.48°\mathrm{C}$

The spread across definitions ($\approx 0.6°\mathrm{C}$) is smaller than, but comparable in order of magnitude to, the parameter-uncertainty-driven spreads emphasized throughout this chapter (e.g. Appendix G's $6\times$ range in $\mathrm{VaR}_{0.999}$) — but unlike parameter uncertainty, **this spread does not shrink with more data**: it is a definitional ambiguity, not a sta-

tistical one, and persists even in the (hypothetical) limit of perfectly known (ξ, μ_t, σ_t) for all t.

Practical guidance. The choice among these three definitions is not merely academic: engineering codes and regulatory frameworks have, at various times and in various jurisdictions, implicitly assumed definition (1) (computing "the" design level from current-climate parameters, ignoring future change) while colloquially using language ("100-year flood") that more closely connotes definition (2) or (3). The IPCC and national climate adaptation guidance increasingly recommend definition (3) (design-life exceedance probability) as the most policy-relevant for infrastructure with a known intended lifespan, since it directly answers the question "what is the probability this asset experiences an exceedance during its service life?" — the quantity ultimately relevant for risk-informed design decisions, connecting directly to the VaR/CVaR framework of Chapter 4 if "exceedance during the design life" is itself treated as the adverse event whose probability is to be controlled.

I.3 Non-Stationary Risk Measures and Their Aggregation

Overview

Chapter 4's risk measures (VaR, CVaR, coherent measures) were defined for a *single* loss random variable L. Under non-stationarity, the loss distribution itself is time-indexed, $L_t \sim F_{L,t}$, raising the question: how should a risk measure be computed and reported for a non-stationary loss process, particularly when the risk measure is to inform a decision (capital allocation, design specification) that must be made *now* but pertains to risks unfolding over a future horizon?

Time-varying VaR and the "current" vs. "horizon-averaged" distinction. Two natural candidates:

$$\mathrm{VaR}_q^{\text{current}} = F_{L,t_0}^{-1}(q) \qquad \text{vs.} \qquad \mathrm{VaR}_q^{\text{horizon}} = F_{\bar{L}}^{-1}(q), \quad \bar{L} \sim \frac{1}{L}\sum_{t=t_0}^{t_0+L-1} F_{L,t} \text{ (mixture)}.$$

$\mathrm{VaR}_q^{\text{current}}$ uses only the distribution at the reference time t_0 — analogous to return-period definition (1) above. $\mathrm{VaR}_q^{\text{horizon}}$ treats the loss over the design life as a draw from a *mixture* of the year-by-year distributions (each year equally likely to be "the" year in which the relevant loss is realized) — analogous in spirit to definition (3), though the mixture formulation is a simplification (it does not capture the *sequential*, "first exceedance" structure of the design-life definition; it instead asks "if a single loss is drawn from a randomly chosen year in the horizon, what is its VaR").

Example — Worked example: capital requirement under a trending severity distribution

Suppose operational losses (Appendix G's setting) exhibit a trend: the GPD scale parameter $\beta_t = \beta_0(1 + \gamma t)$ for some growth rate $\gamma > 0$ (e.g. reflecting growth in transaction volumes, itself increasing the scale of potential fraud losses), with ξ constant. Take $\hat{\xi} = 0.4$, $\beta_0 = \$500000$, $\gamma = 0.05$/year (5% annual growth in scale), and a 10-year capital planning horizon $t = 0, \ldots, 9$.
$\mathrm{VaR}_{0.999}^{\text{current}}$ $(t_0 = 0)$**:** using $\beta_0 = 500000$ directly in the return-level-type formula (Section 2.5, adapted), suppose this gives $\mathrm{VaR}_{0.999}^{\text{current}} \approx \$22\text{M}$ (illustrative; exact value de-

pends on frequency parameters as in Appendix G).

$\text{VaR}_{0.999}^{\text{horizon}}$**:** the mixture over $t = 0, \ldots, 9$ includes years with β_t ranging from $\beta_0 = 500000$ (at $t = 0$) to $\beta_9 = 500000 \times (1 + 0.05 \times 9) = 500000 \times 1.45 = 725000$ (at $t = 9$), a 45% increase in scale. Since $\text{VaR}_{0.999}$ scales *faster* than linearly in β for $\xi > 0$ (the return-level formula $x_T \propto \beta$ exactly, for fixed ξ, λ, T — actually *linearly* in β, per Section 2.5's formula $x_T = u + \frac{\beta}{\xi}[(\lambda T)^{\xi} - 1]$, so a 45% increase in β translates to a 45% increase in the $(x_T - u)$ component), the mixture's 99.9% quantile is influenced disproportionately by the later (higher-β), heavier-tailed years — the mixture distribution's tail is, to leading order, the tail of its *heaviest* component (a finite-mixture analogue of the single-big-jump principle, Section 3.3: the tail of a finite mixture of subexponential distributions is asymptotically dominated by the heaviest-tailed/largest-scale component). This might give $\text{VaR}_{0.999}^{\text{horizon}} \approx \$29\text{M}$ — closer to the $t = 9$ value than to a simple average of the $t = 0$ and $t = 9$ values, illustrating the dominance of the heaviest-tailed component.

Capital implications. If capital is set using $\text{VaR}_{0.999}^{\text{current}} \approx \$22\text{M}$ and held fixed for the 10-year horizon, it will be *under*-capitalized relative to the actual (growing) risk in later years of the horizon — by year 9, the "true" $\text{VaR}_{0.999}$ (using β_9) would itself be approximately $\$22\text{M} \times 1.45 \approx \$31.9\text{M}$, even higher than the horizon-mixture value of $\$29\text{M}$ (since the year-9-only VaR is more extreme than the mixture, which averages in the lower-β early years). **A capital framework based on $\text{VaR}_{0.999}^{\text{current}}$, recalibrated only periodically (e.g. annually), will systematically lag a genuinely growing risk profile** — the lag being most severe immediately *before* each recalibration, when the gap between current capital (based on stale parameters) and current true risk (based on the latest β_t) is largest. This is a structural feature of any periodically-recalibrated risk framework applied to a non-stationary process, distinct from — and additive to — the parameter-estimation uncertainty (Appendix G) and model uncertainty (threshold choice, frequency model) discussed previously.

Synthesis: non-stationarity as a further layer of irreducible uncertainty. This appendix has shown that even *given* a correctly-specified non-stationary model (known trend parameters, known functional form for μ_t, σ_t, β_t), the translation of that model into a single "risk number" (a return level, a VaR) is not unique — it depends on a definitional choice (Section I.2's three return-period concepts; Section I.3's current-vs-horizon VaR) that has no purely statistical resolution. Combined with the within-model parameter uncertainty (Appendix A, D, G) and the between-model uncertainty (is the trend linear? does ξ also vary? Section I.1's discussion), non-stationary extreme value analysis confronts *three nested layers* of uncertainty — parameter, model-specification, and definitional — each of which can individually produce factor-of-2 or larger swings in the final risk number. **The mathematics developed in this volume provides the tools to make each of these layers explicit and to quantify (where quantifiable) their individual and combined contributions** — but no amount of additional mathematical sophistication can, by itself, resolve the definitional layer (Section I.2), which is ultimately a question of *what risk-relevant question is being asked*, to be answered by the decision-maker and the context of the decision, not by the statistician or mathematician alone. This is, in microcosm, the central message of the entire volume: extreme value theory and coherent risk measurement are indispensable tools for *structuring* difficult questions about rare events under uncertainty, but they do not — and cannot — make those questions easy.

Appendix J

Copula Estimation, Tail Dependence Inference, and Multivariate Risk Aggregation

A rigorous treatment of estimating dependence structures for joint extremes, extending Section 3.5 from definitions to working methodology, with full derivations and an integrated worked example in portfolio risk aggregation.

J.1 Parametric Copula Families and Their Tail Behaviour

Overview

Section 3.5 introduced the copula concept and the tail dependence coefficient λ_U, contrasting the (tail-independent) Gaussian copula with the (tail-dependent) Student-t copula. This section surveys the broader landscape of parametric copula families used in practice, derives λ_U for each, and — critically — distinguishes *upper* from *lower* tail dependence, a distinction of central importance for loss data (where the upper tail represents large losses) but often conflated in introductory treatments.

The Archimedean copula family. A copula C is *Archimedean* if it can be written

$$C(u, v) = \psi^{-1}(\psi(u) + \psi(v))$$

for some *generator* $\psi : [0, 1] \to [0, \infty]$, convex, decreasing, $\psi(1) = 0$. Archimedean copulas are popular because they reduce a bivariate (or, with appropriate generalization, multivariate) dependence structure to a single function of one variable, and many standard families arise from simple generator choices.

Definition (Common Archimedean copulas and their generators)

- **Clayton:** $\psi(t) = \frac{1}{\theta}(t^{-\theta} - 1)$, $\theta > 0$. $C(u, v) = (u^{-\theta} + v^{-\theta} - 1)^{-1/\theta}$.
- **Gumbel:** $\psi(t) = (-\log t)^{\theta}$, $\theta \geq 1$. $C(u, v) = \exp\{-[(-\log u)^{\theta} + (-\log v)^{\theta}]^{1/\theta}\}$.
- **Frank:** $\psi(t) = -\log \frac{e^{-\theta t} - 1}{e^{-\theta} - 1}$, $\theta \neq 0$. $C(u, v) = -\frac{1}{\theta} \log \left[1 + \frac{(e^{-\theta u} - 1)(e^{-\theta v} - 1)}{e^{-\theta} - 1}\right]$.

Tail dependence of Archimedean copulas. A general result connects the tail dependence coefficients to the generator's behaviour near 0 (for upper tail) and the behaviour of ψ^{-1} near ∞:

$$\lambda_U = 2 - \lim_{t \to 0^+} \frac{1 - \psi^{-1}(2t)}{1 - \psi^{-1}(t)} \qquad \text{(for upper tail, suitably stated)},$$

$$\lambda_L = \lim_{t \to \infty} \frac{\psi^{-1}(2t)}{\psi^{-1}(t)} \qquad \text{(for lower tail)}.$$

Example — Computing λ_U, λ_L for Clayton and Gumbel

Clayton. $\psi^{-1}(t) = (1 + \theta t)^{-1/\theta}$. For the lower tail: $\psi^{-1}(2t)/\psi^{-1}(t) = \left(\frac{1+2\theta t}{1+\theta t}\right)^{-1/\theta} \to 2^{-1/\theta}$ as $t \to \infty$. So $\lambda_L = 2^{-1/\theta} > 0$ for any $\theta > 0$. For the upper tail, the corresponding limit gives $\lambda_U = 0$. **The Clayton copula has lower tail dependence but is upper-tail independent.**
Gumbel. $\psi^{-1}(t) = \exp(-t^{1/\theta})$. For the upper tail: $1 - \psi^{-1}(2t) \approx (2t)^{1/\theta}/1$ (for small t, using $1 - e^{-x} \approx x$) and similarly $1 - \psi^{-1}(t) \approx t^{1/\theta}$, so the ratio $\to (2t)^{1/\theta}/t^{1/\theta} = 2^{1/\theta}$,

giving $\lambda_U = 2 - 2^{1/\theta}$. For $\theta = 1$ (independence copula, $C(u,v) = uv$), $\lambda_U = 2 - 2 = 0$, correct (independence has no tail dependence). For $\theta \to \infty$ (approaching the comonotonicity/Fréchet upper bound copula $C(u,v) = \min(u,v)$), $2^{1/\theta} \to 1$, so $\lambda_U \to 1$ (perfect tail dependence), also correct. **The Gumbel copula has upper tail dependence ($\lambda_U = 2 - 2^{1/\theta} \in (0,1)$ for $\theta > 1$) but is lower-tail independent ($\lambda_L = 0$).**

The asymmetry is the point. Clayton and Gumbel are, in a sense, mirror images: Clayton captures "crashes together, recoveries independent" (lower-tail dependence for returns, i.e. upper-tail dependence for *losses* if losses are defined as negative returns — terminology requires care!) while Gumbel captures the reverse. For *loss* data (the convention throughout this volume, Section 4.1 onward, where positive values denote losses), **the Gumbel copula's upper-tail dependence is the relevant property for joint catastrophic losses** — two insured perils, or two assets, whose *large losses* tend to co-occur, are well-modelled (at least qualitatively) by a Gumbel copula on the loss variables, while a Clayton copula on the same variables would imply joint *small*-loss (or no-loss) events are correlated but large losses are independent — the opposite of the typically relevant systemic risk concern (Chapter 16).

J.2 Estimation: Maximum Likelihood, IFM, and Rank-Based Methods

Overview

Given bivariate (or multivariate) data, how does one estimate a copula's parameter(s)? Three approaches, in increasing order of computational simplicity and decreasing (generally) statistical efficiency: full maximum likelihood (joint estimation of margins and copula), the Inference Functions for Margins (IFM) two-step method, and fully non-parametric rank-based (pseudo-likelihood) estimation.

Full ML. Given a parametric model $F_1(\cdot;\boldsymbol{\theta}_1)$, $F_2(\cdot;\boldsymbol{\theta}_2)$ for the margins and $C(\cdot,\cdot;\alpha)$ for the copula, the joint density is $f(x,y) = c(F_1(x;\boldsymbol{\theta}_1), F_2(y;\boldsymbol{\theta}_2);\alpha) f_1(x;\boldsymbol{\theta}_1) f_2(y;\boldsymbol{\theta}_2)$, where $c = \partial^2 C/\partial u \partial v$ is the copula density. Full ML maximizes the joint log-likelihood over $(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \alpha)$ simultaneously — statistically efficient but computationally demanding, especially as the number of margins grows (the joint optimization is over a high-dimensional parameter space).

IFM (two-step). Step 1: estimate each margin's parameters $\hat{\boldsymbol{\theta}}_i$ separately by univariate ML (exactly the GEV/GPD estimation of Sections 1.5, 2.4, treating each variable's marginal extreme value behaviour independently). Step 2: transform data to (pseudo-)uniform margins via $\hat{U}_i = F_i(X_i;\hat{\boldsymbol{\theta}}_i)$, then estimate α by maximizing $\sum_t \log c(\hat{U}_{1t}, \hat{U}_{2t};\alpha)$.

Theorem (Asymptotic properties of IFM)

Under standard regularity conditions, the IFM estimator $\hat{\alpha}_{\text{IFM}}$ is consistent and asymptotically normal, with asymptotic variance generally *larger* than that of the full MLE $\hat{\alpha}_{\text{ML}}$ (IFM is asymptotically efficient only in the special case where the margins and copula parameters are asymptotically independent, e.g. for certain elliptical copulas with specific marginal families) — but the efficiency loss is typically modest (often a few percent in asymptotic variance) for commonly-used combinations, while the computational savings (separable optimization) are substantial, especially in high dimensions.

Rank-based (pseudo-likelihood) estimation. Replace the parametric margins entirely: define pseudo-observations $\hat{U}_{it} = \frac{1}{n+1}\text{rank}(X_{it})$ (the scaled empirical rank, among the n observations of variable i), and maximize $\sum_t \log c(\hat{U}_{1t}, \hat{U}_{2t}; \alpha)$ over α alone. This method is *margin-free* — robust to marginal misspecification — at the cost of discarding the marginal information entirely (no GEV/GPD parameters are estimated in this step; they would need to be estimated separately if required).

Example — Worked example: Gumbel copula parameter estimation via rank-based method

Suppose bivariate loss data (X_1, X_2) for two business lines, $n = 200$ paired annual observations. Convert to pseudo-observations $(\hat{U}_{1t}, \hat{U}_{2t}) = (\text{rank}(X_{1t})/(n+1), \text{rank}(X_{2t})/(n+1))$.

The Gumbel copula density is $c(u,v;\theta) = \frac{C(u,v;\theta)}{uv} \frac{[(-\log u)(-\log v)]^{\theta-1}}{[(-\log u)^\theta + (-\log v)^\theta]^{2-1/\theta}} \{[(-\log u)^\theta + (-\log v)^\theta]^{1/\theta} + \theta - 1$ — a somewhat unwieldy expression, but standard software (or direct numerical optimization of the log-likelihood $\sum_t \log c(\hat{U}_{1t}, \hat{U}_{2t}; \theta)$ over $\theta \geq 1$) yields a point estimate.

Suppose this gives $\hat{\theta}_{\text{Gumbel}} = 1.45$ (SE 0.18, via the observed information from the pseudo-likelihood, with an adjustment for the rank-based nature of the pseudo-observations — the "naive" likelihood-based SE understates the true SE because it ignores the estimation error in the marginal ranks themselves; a bootstrap is often used in practice for valid SEs).

Implied λ_U: $\lambda_U = 2 - 2^{1/1.45} = 2 - 2^{0.690}$. Compute $2^{0.690} = \exp(0.690 \times \ln 2) = \exp(0.690 \times 0.6931) = \exp(0.478) \approx 1.613$. So $\hat{\lambda}_U \approx 2 - 1.613 = 0.387$.

Comparison to Kendall's τ. For Gumbel, $\tau = 1 - 1/\theta = 1 - 1/1.45 \approx 0.310$. The sample Kendall's τ (a simple, robust, rank-based concordance measure, computable directly without any copula assumption: $\hat{\tau} = \frac{2}{n(n-1)} \sum_{s<t} \text{sign}[(X_{1s} - X_{1t})(X_{2s} - X_{2t})]$) might independently give $\hat{\tau}_{\text{sample}} \approx 0.305$ — close to the Gumbel-implied 0.310, providing an informal goodness-of-fit check: **if the Gumbel-copula-implied τ matches the (model-free) sample τ, this is at least consistent with (though does not prove) the Gumbel form being appropriate**; a large discrepancy would indicate the Gumbel family is a poor fit regardless of the value of θ chosen.

J.3 Goodness-of-Fit Testing for Copulas

Overview

Having fit a parametric copula, how does one assess whether the chosen family is adequate, as opposed to merely having found the best-fitting member of an inadequate family? This section develops the Cramér–von Mises-type goodness-of-fit test based on the empirical copula process, the standard approach in the copula literature, analogous in spirit to the Kolmogorov-Smirnov-based approach for univariate power laws (Section 3.2).

Definition (Empirical copula and the goodness-of-fit statistic)

The empirical copula is $C_n(u,v) = \frac{1}{n}\sum_{t=1}^{n} \mathbf{1}\{\hat{U}_{1t} \leq u, \hat{U}_{2t} \leq v\}$ (using the pseudo-observations of Section J.2). The Cramér–von Mises statistic is

$$S_n = n \int_{[0,1]^2} [C_n(u,v) - C_{\hat{\alpha}}(u,v)]^2 \, dC_n(u,v),$$

where $C_{\hat{\alpha}}$ is the fitted parametric copula. S_n is computed in practice as a sum over the n pseudo-observation pairs (the integral against dC_n reduces to a finite sum, since C_n is a step function).

Parametric bootstrap for the null distribution. The asymptotic distribution of S_n under H_0 (the fitted family is correct) depends on $\hat{\alpha}$ and is not standard — a parametric bootstrap is used: simulate B datasets from $C_{\hat{\alpha}}$ (with n observations each), re-estimate $\hat{\alpha}^{(b)}$ and compute $S_n^{(b)}$ for each, and obtain a p-value as the proportion of $S_n^{(b)}$ exceeding the observed S_n.

Example — Worked example: comparing Gumbel and Gaussian fits

Continuing Section J.2's example ($\hat{\theta}_{\text{Gumbel}} = 1.45$): suppose $S_n^{\text{Gumbel}} = 0.082$, and a parametric bootstrap with $B = 1000$ gives $p \approx 0.41$ — the Gumbel copula is **not rejected** (no evidence of lack of fit).

For comparison, fit a Gaussian copula to the same data: $\hat{\rho}_{\text{Gauss}} \approx 0.43$ (via, e.g., matching Kendall's τ: for Gaussian copula, $\tau = \frac{2}{\pi}\arcsin\rho$, so $\rho = \sin(\pi\tau/2) = \sin(\pi \times 0.305/2) = \sin(0.479) \approx 0.461$ — close to but not identical to a direct ML estimate, illustrating that different estimation criteria can give somewhat different $\hat{\rho}$ even for the same data). Suppose $S_n^{\text{Gauss}} = 0.156$, with bootstrap $p \approx 0.07$ — **marginal evidence against the Gaussian copula** at conventional significance levels.

The decisive comparison: tail dependence. Even though both copulas might pass (or nearly pass) the overall goodness-of-fit test — S_n is a global measure, averaging fit quality across the entire unit square, including the bulk where both copulas may fit reasonably — the Gaussian copula implies $\lambda_U = 0$ (Section 3.5) while the Gumbel implies $\hat{\lambda}_U \approx 0.387$ (Section J.2). **For risk management purposes, this difference (zero vs. $\approx 39\%$ conditional probability of joint extreme losses) is far more consequential than the overall goodness-of-fit p-values might suggest**, echoing Section 3.5's warning about the Gaussian copula: a copula can pass an *overall* goodness-of-fit test (because the test's power is dominated by the bulk of the data, where most of the n observations lie) while being severely misspecified in the *tail* (where few or no observations directly inform the fit) — precisely the region of interest for risk aggregation. **A dedicated tail-focused goodness-of-fit assessment** — e.g. comparing $C_n(u,v)$ to $C_{\hat{\alpha}}(u,v)$ restricted to $u, v > 0.9$ or similar — is recommended whenever the downstream application is risk-focused, supplementing (not replacing) the global S_n test.

J.4 Integrated Worked Example: Two-Asset Portfolio Tail Risk

Problem statement

A portfolio consists of two assets with annual loss distributions $L_1 \sim \text{GPD}(\xi_1 = 0.3, \beta_1 = 2)$ (in \$M, above respective thresholds $u_1 = u_2 = 5$) and $L_2 \sim \text{GPD}(\xi_2 = 0.25, \beta_2 = 1.5)$, dependent via a Gumbel copula with $\theta = 1.45$ (as estimated in Section J.2). Compute $\text{CVaR}_{0.99}$ for the combined portfolio loss $L_1 + L_2$, and compare to the sum of individual $\text{CVaR}_{0.99}$'s (i.e. compute the diversification ratio of Appendix D.2).

Step 1: Individual CVaRs. Using Section 4.2's formula, $\text{CVaR}_{0.99}(L_i) = \text{VaR}_{0.99}(L_i) + e_i(\text{VaR}_{0.99}(L_i))$. We need $\text{VaR}_{0.99}$ for each — suppose, given the GPD parameters and an (unspecified for brevity, but analogous to earlier worked examples) exceedance rate, this yields $\text{VaR}_{0.99}(L_1) \approx \$18.2\text{M}$ and $\text{VaR}_{0.99}(L_2) \approx \$14.6\text{M}$.

Mean excess: $e_1(18.2) = \frac{\beta_1 + \xi_1(18.2 - u_1)}{1 - \xi_1} = \frac{2 + 0.3 \times 13.2}{0.7} = \frac{2 + 3.96}{0.7} = \frac{5.96}{0.7} \approx 8.51$. So $\text{CVaR}_{0.99}(L_1) \approx 18.2 + 8.51 \approx 26.71$.

$e_2(14.6) = \frac{1.5 + 0.25 \times 9.6}{0.75} = \frac{1.5 + 2.4}{0.75} = \frac{3.9}{0.75} = 5.2$. So $\text{CVaR}_{0.99}(L_2) \approx 14.6 + 5.2 \approx 19.8$.

Sum of individual CVaRs: $26.71 + 19.8 = 46.51$.

Step 2: Joint simulation under the Gumbel copula. To compute $\text{CVaR}_{0.99}(L_1 + L_2)$, simulate $M = 10^6$ pairs $(U_1, U_2) \sim C_{\text{Gumbel}}(\theta = 1.45)$ (via, e.g., the Marshall-Olkin algorithm for Archimedean copulas, which represents (U_1, U_2) via a shared positive stable random variable S with $U_i = \psi(-\log V_i / S)$ for independent uniforms V_i — the details of this simulation algorithm are standard but we omit them, focusing on the resulting risk calculation), then $L_i = u_i + \hat{H}^{-1}_{\xi_i, \beta_i}(U_i)$ (inverting the GPD CDF), and $S = L_1 + L_2$.

Step 3: Results. Suppose the simulation yields $\text{VaR}_{0.99}(S) \approx \$29.8\text{M}$ and $\text{CVaR}_{0.99}(S) \approx \$41.3\text{M}$.

Diversification ratio: $DR = \frac{\text{CVaR}_{0.99}(L_1) + \text{CVaR}_{0.99}(L_2)}{\text{CVaR}_{0.99}(S)} = \frac{46.51}{41.3} \approx 1.126$.

Interpretation. A diversification ratio of 1.126 indicates a modest ($\approx 11\%$) reduction in CVaR from combining the two assets, compared to holding them as entirely separate risk buckets (i.e. $DR = 1$ would mean no benefit). Recall from Appendix D.2 that for two i.i.d. *independent* heavy-tailed risks, $DR \to 1$ as the confidence level $\to 1$ (single-big-jump). Here, with $\hat{\lambda}_U \approx 0.387$ (substantial but not perfect tail dependence) and $\xi_1 \neq \xi_2$ (somewhat different tail heaviness), $DR \approx 1.126$ at the 99% level represents an intermediate case — meaningfully different from both $DR = 1$ (no benefit, as under perfect tail dependence or extreme confidence levels for i.i.d. heavy tails) and the $DR \approx 1.4$ that might obtain for genuinely tail-independent ($\lambda_U = 0$), similarly-scaled risks at this confidence level.

Step 4: Sensitivity to the copula choice. Re-running Step 2–3 with a Gaussian copula calibrated to the *same* Kendall's $\tau \approx 0.305$ (i.e. $\hat{\rho}_{\text{Gauss}} \approx 0.461$, from Section J.3) — which, recall, implies $\lambda_U = 0$ despite the substantial τ — might yield $\text{CVaR}^{\text{Gauss}}_{0.99}(S) \approx \$38.9\text{M}$, giving $DR^{\text{Gauss}} = 46.51/38.9 \approx 1.196$.

The Gumbel-vs-Gaussian copula choice changes $\text{CVaR}_{0.99}(S)$ from $\$41.3$M to $\$38.9$M — a $\$2.4$M ($\approx$6%) difference — and the diversification ratio from 1.126 to 1.196, despite both copulas having (by construction) the *same* Kendall's τ, i.e. the same "overall" dependence as measured by this standard rank-correlation statistic. The entire 6% difference in the risk number is attributable to the *tail* dependence structure ($\lambda_U \approx 0.387$ vs. $\lambda_U = 0$) — a feature invisible to Kendall's τ (which is a property of the *entire* copula, weighted toward the bulk where most probability mass lies) but, per Section J.3's discussion, potentially decisive for risk-relevant conclusions.

> **Closing synthesis for Appendix J.** This appendix has traced a complete path from the abstract tail dependence concepts of Section 3.5 to a concrete portfolio risk number: parametric copula families and their (often asymmetric — Section J.1) tail dependence properties; three estimation approaches of varying robustness and efficiency (Section J.2); the subtlety that global goodness-of-fit tests may not detect tail misspecification (Section J.3); and a worked aggregation example quantifying the resulting sensitivity of CVaR and the diversification ratio to the copula choice (Section J.4). The overarching lesson echoes — in the explicitly multivariate setting — the message of Appendix G's univariate operational risk case study: **the final risk number depends on modelling choices (here, the copula family) that are difficult to discriminate between using standard statistical tools, yet which can move the risk number by**

amounts (here, $\approx 6\%$ in CVaR, translating to potentially material differences in required capital for a large portfolio) that are far from negligible for the decisions the risk number is meant to inform. Combined with Appendix I's analysis of non-stationarity and the parameter-uncertainty analyses of Appendices A, D, and G, this appendix completes the volume's portrait of extreme value-based risk quantification as a discipline defined as much by the careful *characterization and communication* of irreducible uncertainties as by the (considerable, and genuinely useful) point-estimation machinery developed in the main chapter.

Appendix K

Bayesian Extreme Value Analysis: Priors, MCMC, and Predictive Inference

A complete development of Bayesian GEV/GPD inference, from prior elicitation through Metropolis–Hastings implementation to posterior predictive risk quantities, extending the sketch of Appendix D.1.

K.1 Prior Elicitation for the Shape Parameter

Overview

The single most consequential modelling choice in Bayesian extreme value analysis is the prior on ξ. Because, as emphasized throughout this volume, ξ is weakly identified by data of realistic size, the prior can materially influence the posterior — in sharp contrast to typical Bayesian analyses of well-identified parameters, where the prior's influence is asymptotically negligible. This section surveys the principal proposals for ξ-priors and their consequences.

The flat (improper) prior and its dangers. $\pi(\xi) \propto 1$ on $\mathbb{R}$ is the naive "non-informative" choice. However, the GEV/GPD likelihood (Sections 1.5, 2.4) is not integrable over all of $\mathbb{R}$ in ξ — for $\xi \to -\infty$, the likelihood can behave pathologically (related to the support-depends-on-parameters issue noted for $\xi < -1/2$ in Sections 1.3, 2.4), and for $\xi \to +\infty$, the likelihood does not decay fast enough to ensure a proper posterior in general. **A flat prior on ξ can yield an improper posterior** — a well-documented pitfall in naive Bayesian EVT implementations.

Definition (Proper priors for ξ in common use)

- **Uniform on a bounded interval:** $\xi \sim \text{Uniform}(-0.5, 0.5)$ or $\text{Uniform}(-1, 1)$ — enforces a physically-motivated bound but introduces a hard truncation that can distort the posterior near the boundary if the likelihood favours values near or beyond it.
- **Beta-type prior on a transformed scale:** e.g. $\frac{\xi+0.5}{1} \sim \text{Beta}(a, b)$ for $\xi \in (-0.5, 0.5)$, allowing the prior to place more mass near $\xi = 0$ (Gumbel) if desired, via $a = b > 1$.
- **Normal prior:** $\xi \sim \mathcal{N}(0, \tau^2)$ for some τ (e.g. $\tau = 0.2$ or $\tau = 0.5$) — proper on all of $\mathbb{R}$, automatically ensuring posterior propriety, with the choice of τ directly controlling how strongly the prior pulls ξ toward 0 (the Gumbel case).
- **Martins–Stedinger prior:** $\pi(\xi) \propto (0.5 + \xi)^{p-1}(0.5 - \xi)^{q-1}$ for $\xi \in (-0.5, 0.5)$, a Beta-type prior on $(-0.5, 0.5)$ specifically motivated by physical considerations in hydrology (the choice $p = q = 6$, for instance, concentrates substantial prior mass near $\xi = 0$ while still allowing $|\xi|$ up to 0.5).

Example — Quantifying prior influence: a sensitivity analysis

Consider the running example of Section 1.5 ($m = 50$ block maxima, $\hat{\xi}_{\text{ML}} = 0.21$, SE $= 0.13$). We compare posteriors under three priors, using a normal approximation to the likelihood (valid for illustration, given $m = 50$ is large enough for the asymptotic normality of Section 1.5 to be a reasonable — if imperfect — approximation to the likelihood shape):

Prior A: $\xi \sim \mathcal{N}(0, 0.5^2)$ **(weakly informative, centered at Gumbel).** The posterior is (approximately, treating both likelihood and prior as Gaussian, so the posterior is also Gaussian by conjugacy) $\mathcal{N}(\xi_{\text{post}}, \sigma^2_{\text{post}})$ with

$$\frac{1}{\sigma^2_{\text{post}}} = \frac{1}{0.13^2} + \frac{1}{0.5^2} = \frac{1}{0.0169} + \frac{1}{0.25} = 59.17 + 4.00 = 63.17,$$

$$\xi_{\text{post}} = \sigma^2_{\text{post}} \left[\frac{0.21}{0.0169} + \frac{0}{0.25}\right] = \frac{1}{63.17} \times 12.43 \approx 0.1968.$$

$$\sigma_{\text{post}} = \sqrt{1/63.17} \approx 0.1258.$$

Posterior mean shifted slightly toward 0 (from 0.21 to 0.197), with SD reduced from 0.13 to 0.126 — modest prior influence.

Prior B: $\xi \sim \mathcal{N}(0, 0.1^2)$ **(more informative, strong pull toward Gumbel).**

$$\frac{1}{\sigma^2_{\text{post}}} = 59.17 + \frac{1}{0.01} = 59.17 + 100 = 159.17, \qquad \xi_{\text{post}} = \frac{1}{159.17} \times 12.43 \approx 0.0781.$$

$$\sigma_{\text{post}} = \sqrt{1/159.17} \approx 0.0793.$$

Posterior mean shifted substantially, from 0.21 to 0.078 — **more than halved**, and now comfortably within the "light tail" region, despite the data's point estimate (0.21) suggesting moderate heaviness.

Prior C: $\xi \sim \text{Uniform}(-0.5, 0.5)$**.** For ξ well within the bounds (as 0.21 and the likelihood's bulk are), this prior is locally flat and the posterior $\approx$ the (renormalized) likelihood itself, i.e. $\xi_{\text{post}} \approx 0.21$, $\sigma_{\text{post}} \approx 0.13$ — essentially unchanged from the MLE.

Consequence for $\text{CVaR}_{0.99}$**.** Recalling Appendix D.1's finding that the posterior *predictive* tail probability is dominated by the upper tail of the ξ-posterior: under Prior C ($\xi_{\text{post}} \approx 0.21 \pm 0.13$), the posterior places substantial mass on $\xi > 0.34$ (one SD above the mean), in the heavy-tail-CVaR regime of Section 4.2's main example. Under Prior B ($\xi_{\text{post}} \approx 0.078 \pm 0.079$), $\xi > 0.34$ is over 3 SDs from the posterior mean — a tail event under the posterior itself, contributing negligibly to the predictive average.

The choice between Prior B and Prior C — both "reasonable" on their face — can change the predictive $\text{CVaR}_{0.99}$ **by a large factor**, via exactly the convexity-driven amplification mechanism of Appendix D.1, now operating on the *prior-to-posterior* mapping rather than the posterior-to-predictive mapping, but with the same qualitative consequence: **seemingly minor, defensible choices in the analysis can dominate the final risk number**, a theme this volume has now illustrated at the level of estimation method (ML vs. PWM, Section 2.4), threshold choice (Section 2.3, Appendix G), copula family (Appendix J), non-stationarity specification (Appendix I), and now prior specification.

K.2 Metropolis–Hastings for the GPD Posterior

Overview

Section K.1's normal-approximation analysis is illustrative but not exact — the true GEV/GPD posterior is generally not Gaussian, particularly for small samples or ξ near boundary values. This section develops the Metropolis–Hastings (MH) algorithm for sampling the exact GPD posterior, including the practical details (proposal tuning,

reparametrization) that determine whether the algorithm performs well in practice.

The target posterior. For GPD exceedances $Y_1, \dots, Y_k$ and prior $\pi(\xi, \beta) = \pi(\xi)\pi(\beta)$ (often $\pi(\beta) \propto 1/\beta$ for $\beta > 0$, the Jeffreys-type scale-invariant prior), the posterior is

$$\pi(\xi, \beta \mid \mathbf{y}) \propto \left[\prod_{i=1}^{k} h_{\xi,\beta}(y_i)\right] \pi(\xi)\pi(\beta) = \beta^{-k} \prod_{i=1}^{k} \left(1 + \xi \frac{y_i}{\beta}\right)^{-1/\xi - 1} \pi(\xi)\pi(\beta),$$

subject to the support constraint $1 + \xi y_i/\beta > 0$ for all i.

Algorithm (Random-walk Metropolis–Hastings for (ξ, β))

1. Initialize $(\xi^{(0)}, \beta^{(0)})$ at, e.g., the MLE (Section 2.4).
2. For $j = 1, \dots, J$:
   (a) Propose $(\xi', \beta') = (\xi^{(j-1)} + \epsilon_\xi, \beta^{(j-1)} \cdot e^{\epsilon_{\log\beta}})$ where $\epsilon_\xi \sim \mathcal{N}(0, \tau_\xi^2)$, $\epsilon_{\log\beta} \sim \mathcal{N}(0, \tau_\beta^2)$ (random-walk proposal on ξ directly, but on $\log\beta$ to respect positivity — a standard reparametrization).
   (b) If $\beta' \le 0$ or the support constraint fails for any y_i under (ξ', β'), reject (set $(\xi^{(j)}, \beta^{(j)}) = (\xi^{(j-1)}, \beta^{(j-1)})$).
   (c) Otherwise compute the acceptance ratio

$$r = \frac{\pi(\xi', \beta' \mid \mathbf{y})}{\pi(\xi^{(j-1)}, \beta^{(j-1)} \mid \mathbf{y})} \times \frac{\beta'}{\beta^{(j-1)}}$$

   (the extra factor $\beta'/\beta^{(j-1)}$ is the Jacobian of the $\log\beta$ reparametrization, required for detailed balance on the original (ξ, β) scale).
   (d) Accept (ξ', β') with probability $\min(1, r)$; else retain $(\xi^{(j-1)}, \beta^{(j-1)})$.

Tuning the proposal scales τ_ξ, τ_β. The standard target is an acceptance rate of roughly 20–40% for random-walk MH in 2 dimensions (the well-known "optimal" rate, derived from diffusion-limit theory, is closer to 23.4% for high-dimensional Gaussian targets, but 2D problems tolerate a somewhat wider range). Proposal scales are typically tuned via a pilot run: start with $\tau_\xi = 0.5 \times \mathrm{SE}(\hat{\xi}_{\mathrm{ML}})$, $\tau_{\log\beta} = 0.5 \times \mathrm{SE}(\hat{\beta}_{\mathrm{ML}})/\hat{\beta}_{\mathrm{ML}}$ (the latter approximating the SE of $\log\beta$ via the delta method, $\mathrm{SE}(\log\beta) \approx \mathrm{SE}(\beta)/\beta$), and adjust based on the observed acceptance rate.

Example — Numerical MCMC trace and diagnostics (illustrative)

For the Section 2.4 worked example ($k = 180$ exceedances, $\hat{\xi}_{\mathrm{ML}} = 0.34$, $\hat{\beta}_{\mathrm{ML}} = 9.8$), suppose an MH chain of $J = 50000$ iterations (after 5000 burn-in) with $\tau_\xi = 0.05$, $\tau_{\log\beta} = 0.08$, prior $\xi \sim \mathcal{N}(0, 0.5^2)$, $\pi(\beta) \propto 1/\beta$, achieves an acceptance rate of 0.31 (within the target range).
Posterior summaries: posterior mean $\bar{\xi} \approx 0.338$, posterior SD ≈ 0.091 (compare to $\mathrm{SE}(\hat{\xi}_{\mathrm{ML}}) = 0.094$ from Section 2.4 — close, as expected for a moderately informative prior with $k = 180$ providing substantial likelihood information, so the prior's influence here is modest, unlike the small-m example of Section K.1). Posterior 95% credible interval for ξ: $(0.165, 0.524)$ — note this interval *includes* 0.5 (the infinite-variance boundary), with posterior probability $P(\xi > 0.5 \mid \mathbf{y}) \approx 0.031$ — a directly interpretable Bayesian statement ("a 3.1% posterior probability that the severity distribution has infinite variance") that has no precise frequentist analogue (a frequentist con-

fidence interval does not admit a probability statement about the parameter itself).
Convergence diagnostics. The Gelman–Rubin $\hat{R}$ statistic (comparing within-chain to between-chain variance across multiple independent chains started from dispersed initial values) should be < 1.01 for all parameters; the effective sample size (accounting for autocorrelation in the MH chain, which is typically substantial for random-walk proposals — autocorrelation times of 10–50 iterations are common) might be $\text{ESS}(\xi) \approx 1800$ out of 50000 raw iterations, i.e. an efficiency of $\approx 3.6\%$, typical for random-walk MH on this type of target.

K.3 Posterior Predictive Risk Measures

Overview

Section D.1 introduced the posterior predictive distribution conceptually. This section makes the computation fully explicit for VaR and CVaR, using the MCMC output of Section K.2, and discusses the important distinction between the *posterior predictive VaR* and the *Bayesian point estimate of VaR* — two related but distinct quantities often conflated in applied work.

Definition (Posterior predictive quantile vs. posterior mean of the quantile)

Posterior predictive VaR, VaR_q^{PP}: the q-quantile of the posterior predictive distribution,

$$\text{VaR}_q^{\text{PP}} = \left(F^{\text{PP}}\right)^{-1}(q), \qquad F^{\text{PP}}(y) = \int H_{\xi,\beta}(y)\,\pi(\xi,\beta \mid \mathbf{y})\,d\xi\,d\beta.$$

Posterior mean of VaR, $\mathbb{E}[\text{VaR}_q \mid \mathbf{y}]$: the posterior expectation of the (parameter-dependent) quantile function,

$$\mathbb{E}[\text{VaR}_q(\xi,\beta) \mid \mathbf{y}] = \int \left(H_{\xi,\beta}\right)^{-1}(q)\,\pi(\xi,\beta \mid \mathbf{y})\,d\xi\,d\beta.$$

These are generally different quantities. VaR_q^{PP} is the quantile of a *mixture* (over the posterior) of GPDs; $\mathbb{E}[\text{VaR}_q \mid \mathbf{y}]$ is the *mean* (over the posterior) of the quantile of each individual GPD. Since the quantile function $H^{-1}_{\xi,\beta}(q)$ is a *convex* function of ξ for fixed q close to 1 (the same convexity driving Appendix D.1's and Section K.1's results), and since for a mixture, the q-quantile of the mixture is generally *less* than the q-quantile-weighted-average in the presence of such convexity (a manifestation of the general fact that quantiles of mixtures do not equal mixtures of quantiles, except in special cases), we expect $\text{VaR}_q^{\text{PP}} \lesssim \mathbb{E}[\text{VaR}_q \mid \mathbf{y}]$ in typical cases — though the relationship is not a strict inequality in full generality and depends on the shape of the posterior.

Example — Numerical comparison using the MCMC output

Using the 45000 post-burn-in MCMC draws $(\xi^{(j)}, \beta^{(j)})$ from Section K.2:
$\mathbb{E}[\text{VaR}_{0.99} \mid \mathbf{y}]$**:** for each draw, compute $\text{VaR}_{0.99}(\xi^{(j)}, \beta^{(j)}) = u + \frac{\beta^{(j)}}{\xi^{(j)}}[(\lambda T)^{\xi^{(j)}} - 1]$ (Section 2.5's formula, with $u = 30, \lambda = 4.5, T = 100$ as in the running example), then average over j. This is a direct Monte Carlo average over the MCMC output — straightforward. Suppose this gives $\mathbb{E}[\text{VaR}_{0.99} \mid \mathbf{y}] \approx 237.8$ (compare to the plug-in $\text{VaR}_{0.99} \approx 231.1$ from Section 2.5, using the MLE point estimates — the Bayesian posterior-mean es-

timate is somewhat higher, again reflecting the convexity/Jensen effect, though the shift here, 237.8 vs. 231.1, a $\approx 3\%$ increase, is more modest than some of Appendix D.1's more extreme illustrations, reflecting the relatively well-constrained posterior here with $k = 180$).

$\text{VaR}^{\text{PP}}_{0.99}$**:** requires, for each draw $(\xi^{(j)}, \beta^{(j)})$, evaluating the GPD survival function $\overline{H}_{\xi^{(j)},\beta^{(j)}}$ at a grid of y-values, averaging across j to get $\overline{F}^{\text{PP}}(y)$, then finding y such that $\overline{F}^{\text{PP}}(y) = 0.01$ (in the appropriate annualized sense, matching Section 2.5's setup). Suppose this gives $\text{VaR}^{\text{PP}}_{0.99} \approx 233.5$.

Both Bayesian quantities (237.8 **and** 233.5**) exceed the plug-in MLE-based value** (231.1), consistent with the general expectation that ignoring parameter uncertainty (plug-in) understates risk (Appendix D.1's central finding), with $\mathbb{E}[\text{VaR}_{0.99} \mid \mathbf{y}] > \text{VaR}^{\text{PP}}_{0.99}$ as anticipated from the Jensen-type argument above, though the gap between the two Bayesian quantities (237.8 vs 233.5, $\approx 1.8\%$) is here smaller than the gap between either and the plug-in value ($\approx 2 - 3\%$) — for this moderately-sized dataset ($k = 180$), the choice *between* the two Bayesian summaries is less consequential than the choice of *whether to go Bayesian at all* (i.e. whether to account for parameter uncertainty), though Appendix K.1's small-sample example ($m = 50$) would likely show a larger gap between $\mathbb{E}[\text{VaR} \mid \mathbf{y}]$ and VaR^{PP}, since the posterior is wider and the convexity effects more pronounced.

Recommendation. For risk-reporting purposes, VaR^{PP}_q (the posterior predictive quantile) is generally the more directly interpretable and decision-relevant quantity: it answers the question "what level will next year's loss, accounting for everything we don't know about (ξ, β), exceed with probability $1 - q$?" — precisely the question VaR_q is meant to answer, now correctly incorporating parameter uncertainty into the predictive distribution for the *observable* (a future loss), rather than treating VaR_q as a property of the (unobservable, uncertain) *parameters* and averaging over that uncertainty (which is what $\mathbb{E}[\text{VaR}_q \mid \mathbf{y}]$ does). This distinction — between uncertainty about an observable future quantity (the predictive distribution) and uncertainty about an unobservable parameter (the posterior distribution of a function of the parameter) — is a recurring subtlety in Bayesian risk analysis, and its resolution in favour of the predictive-distribution interpretation is, in the authors' assessment, the conceptually correct one for essentially all risk-management applications, even though $\mathbb{E}[\cdot \mid \mathbf{y}]$-type quantities are sometimes easier to compute (as a simple posterior average, Section K.3's first calculation) and hence more commonly reported in practice — a gap between ease of computation and conceptual correctness that closes this appendix's, and this volume's, recurring theme of the distance between mathematically convenient point estimates and the fuller uncertainty quantification that rigorous risk analysis demands.

Appendix L

Mathematical Prerequisites

A concise treatment of measure theory, integration, characteristic functions, and functional analysis as needed for the proofs and derivations of Chapter 1. Readers fluent in these areas may skim or skip; readers less familiar will find here the specific results used, stated in the forms in which they are applied.

L.1 Probability Spaces and Measurability

Definition (Probability space)

A *probability space* is a triple $(\Omega, \mathcal{F}, P)$ where Ω is a non-empty set (the *sample space*), $\mathcal{F}$ is a σ-algebra on Ω (a collection of subsets closed under complement and countable union, with $\emptyset, \Omega \in \mathcal{F}$), and $P : \mathcal{F} \to [0,1]$ is a probability measure ($P(\Omega) = 1$, P countably additive on disjoint sets).

Definition (Random variable and distribution)

A function $X : \Omega \to \mathbb{R}$ is a *random variable* if it is $\mathcal{F}/\mathcal{B}(\mathbb{R})$-measurable ($\mathcal{B}(\mathbb{R})$ the Borel σ-algebra on $\mathbb{R}$), i.e. $\{X \le x\} \in \mathcal{F}$ for all $x \in \mathbb{R}$. The *distribution* (law) of X is the probability measure μ_X on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ defined by $\mu_X(B) = P(X \in B)$ for $B \in \mathcal{B}(\mathbb{R})$. The *cumulative distribution function* (CDF) is $F_X(x) = P(X \le x) = \mu_X((-\infty, x])$.

CDF properties. Any CDF $F : \mathbb{R} \to [0,1]$ satisfies: (i) F is right-continuous; (ii) $\lim_{x\to-\infty} F(x) = 0$, $\lim_{x\to\infty} F(x) = 1$; (iii) F is non-decreasing. Conversely, any function satisfying (i)–(iii) is the CDF of some probability measure on $\mathbb{R}$ (this is Lebesgue–Stieltjes integration theory).

Quantile function. The *generalized inverse* $F^{-1}(u) = \inf\{x : F(x) \ge u\}$ for $u \in (0,1)$ is the quantile function (Section 1.3). It satisfies: $F^{-1}(u) \le x \iff u \le F(x)$ (for all $u \in (0,1)$ and x at which F is continuous); F^{-1} is left-continuous and non-decreasing; and if $U \sim \text{Uniform}(0,1)$, then $F^{-1}(U) \sim F$ (the probability integral transform, used extensively in copula simulation, Appendix J.2).

L.2 Convergence Concepts

Definition (Convergence modes)

Let $X, X_1, X_2, \ldots$ be random variables on $(\Omega, \mathcal{F}, P)$.

- *Almost sure* (*a.s.*) *convergence*: $X_n \to X$ a.s. if $P(\lim_{n\to\infty} X_n = X) = 1$.
- *Convergence in probability*: $X_n \xrightarrow{P} X$ if $P(|X_n - X| > \epsilon) \to 0$ for all $\epsilon > 0$.
- *Convergence in L^p*: $X_n \xrightarrow{L^p} X$ if $\mathbb{E}[|X_n - X|^p] \to 0$ for $p \ge 1$.
- *Convergence in distribution* (*weak convergence*): $X_n \xrightarrow{d} X$ if $F_{X_n}(x) \to F_X(x)$ at all continuity points x of F_X.

The implications a.s. $\Rightarrow$ in probability $\Rightarrow$ in distribution hold; $L^p \Rightarrow$ in probability. Convergence in distribution does *not* imply convergence in probability in general (it is the weakest mode), yet is the mode appearing throughout Chapters 1–2 (the FTG theorem, the PBdH theorem, the CLT/GCLT of Section 3.4 are all statements of con-

vergence in distribution).

Theorem (Portmanteau theorem for weak convergence)

$X_n \xrightarrow{d} X$ if and only if any one of the following holds: (i) $\mathbb{E}[f(X_n)] \to \mathbb{E}[f(X)]$ for all bounded continuous $f : \mathbb{R} \to \mathbb{R}$; (ii) $\limsup_n P(X_n \in F) \le P(X \in F)$ for all closed sets F; (iii) $\liminf_n P(X_n \in G) \ge P(X \in G)$ for all open sets G; (iv) $P(X_n \in B) \to P(X \in B)$ for all Borel sets B with $P(X \in \partial B) = 0$.

The portmanteau theorem is used implicitly throughout Chapter 1 whenever the FTG limit $F^n(a_n x + b_n) \to G(x)$ is verified at continuity points of G — this is exactly convergence in distribution (condition (iv) with $B = (-\infty, x]$, whose boundary $\partial B = \{x\}$ has $P(X = x) = 0$ whenever G is continuous at x).

Uniform integrability and convergence of moments. If $X_n \xrightarrow{d} X$ and the sequence $\{|X_n|^p\}$ is uniformly integrable (i.e. $\lim_{M\to\infty} \sup_n \mathbb{E}[|X_n|^p \mathbf{1}\{|X_n| > M\}] = 0$), then $\mathbb{E}[|X_n|^p] \to \mathbb{E}[|X|^p]$. This condition is used when going from "convergence in distribution of normalized maxima" to "convergence of moments of normalized maxima" — a non-trivial step requiring the uniform integrability condition on the sequence $M_n/a_n - b_n/a_n$, which fails in general (and explains why, e.g., the mean of M_n properly normalized does not always converge to $\mathbb{E}[G]$, depending on regularity of the tail).

L.3 Characteristic Functions and the Continuity Theorem

Definition (Characteristic function)

The *characteristic function* of a random variable X is $\varphi_X(t) = \mathbb{E}[e^{itX}] = \int e^{itx}\, dF_X(x)$ for $t \in \mathbb{R}$. Properties: $\varphi_X(0) = 1$; $|\varphi_X(t)| \le 1$; φ_X is uniformly continuous. If $\mathbb{E}[|X|^k] < \infty$, then φ_X is k-times differentiable and $\varphi_X^{(k)}(0) = i^k \mathbb{E}[X^k]$.

Theorem (Lévy's continuity theorem)

$X_n \xrightarrow{d} X$ if and only if $\varphi_{X_n}(t) \to \varphi_X(t)$ for all $t \in \mathbb{R}$.

This theorem underlies the proof of the GCLT (Section 3.4): one shows $\varphi_{S_n/a_n}(t) \to \varphi_{X_\alpha}(t)$ (pointwise in t) for the α-stable characteristic function φ_{X_α}, and invokes continuity theorem to conclude $S_n/a_n \xrightarrow{d} X_\alpha$. The same approach works for the CLT ($\alpha = 2$, Gaussian limit) and underlies the "characteristic function method" for proving the FTG theorem as an alternative to the functional-equation proof of Appendix F.1.

Example — Characteristic function approach to the Gumbel limit

For $M_n = \max(X_1, \dots, X_n)$ with $X_i \sim \text{Exp}(1)$, normalizing constants $a_n = 1, b_n = \log n$, so $\tilde{M}_n = (M_n - \log n) \xrightarrow{d} \Lambda$ (Gumbel). The characteristic function of $\tilde{M}_n$:

$$\varphi_{\tilde{M}_n}(t) = \mathbb{E}[e^{it\tilde{M}_n}] = \mathbb{E}[e^{it(M_n - \log n)}] = n^{-it}\mathbb{E}[e^{itM_n}].$$

$\mathbb{E}[e^{itM_n}] = \int_0^\infty e^{itx} n e^{-x}(1-e^{-x})^{n-1} dx$ (using $f_{M_n}(x) = ne^{-x}(1-e^{-x})^{n-1}$ for i.i.d. Exp(1)). Via the substitution $u = 1 - e^{-x}$ and a Laplace-transform type calculation

(standard but lengthy), this evaluates to $\varphi_{\tilde{M}_n}(t) \to \Gamma(1-it)$ as $n \to \infty$, where Γ is the gamma function with complex argument. The characteristic function of the Gumbel distribution Λ is exactly $\varphi_\Lambda(t) = \Gamma(1-it)$ — confirming the convergence, and yielding (as a bonus) the Gumbel's moment formula $\mathbb{E}[\Lambda^k] = \Gamma^{(k)}(1)/(-i)^k$ at $t = 0$, directly reproducing the $\mathbb{E}[X] = \gamma$ (Euler–Mascheroni constant) result via $\Gamma'(1) = -\gamma$.

L.4 Key Results from Real Analysis

The dominated convergence theorem (DCT). If $X_n \to X$ a.s. and $|X_n| \le Y$ a.s. for all n, with $\mathbb{E}[Y] < \infty$, then $\mathbb{E}[X_n] \to \mathbb{E}[X]$. DCT is used throughout Chapter 1 to exchange limits and expectations: e.g. in showing $\mathbb{E}[M_n/a_n] \to \mathbb{E}[G]$ (under uniform integrability), and in the proof of Karamata's theorem (Section 3.1, Appendix F.1).

Monotone convergence theorem (MCT). If $0 \le X_n \nearrow X$ a.s., then $\mathbb{E}[X_n] \nearrow \mathbb{E}[X]$ (no finiteness assumption on $\mathbb{E}[X]$). Used in showing that the GEV mean exists iff $\xi < 1$ (by expressing $\mathbb{E}[X] = \int_0^\infty P(X > x)dx$ and checking convergence via MCT for the integrand).

Fubini–Tonelli theorem. For a product measure space, if $f \ge 0$ (Tonelli) or $\int\int |f|\, d\mu\, d\nu < \infty$ (Fubini), then iterated integrals may be exchanged: $\int\int f\, d\mu\, d\nu = \int\int f\, d\nu\, d\mu$. Used in the proof that $\mathbb{E}[X^k] = \int_0^\infty kx^{k-1}P(X > x)\, dx$ for $X \ge 0$ (by writing $X^k = \int_0^X kx^{k-1}dx$ and exchanging $\mathbb{E}$ with the inner integral via Fubini).

Fatou's lemma. For non-negative X_n: $\mathbb{E}[\liminf X_n] \le \liminf \mathbb{E}[X_n]$. Provides the direction of inequality when DCT does not apply (e.g. showing that $\mathbb{E}[\xi]$ where ξ is a limit of estimators is at most the limit of expected estimates, even when uniform integrability fails).

L.5 Regular Variation: Technical Supplement

This section collects technical results on slowly/regularly varying functions (Section 3.1) used without full proof in the main text.

Theorem (Representation theorem, detailed statement)

$L \in RV_0$ (slowly varying at ∞) if and only if there exist functions $c(x) \to c \in (0, \infty)$ and $\epsilon(x) \to 0$ as $x \to \infty$ such that

$$L(x) = \exp\left(\int_a^x \frac{\epsilon(t)}{t}\, dt\right) \cdot c(x)$$

for some $a > 0$ and all $x \ge a$. In the normalized case (Karamata's canonical form), $c(x) \equiv c$ (constant) and $L(x) = c \exp(\int_a^x \epsilon(t)/t\, dt)$.

Theorem (Uniform convergence on compact sets)

If $L \in RV_0$, then for any $0 < a < b < \infty$:

$$\sup_{t \in [a,b]} \left| \frac{L(tx)}{L(x)} - 1 \right| \to 0 \quad \text{as } x \to \infty.$$

Theorem (Characterization of RV_ρ via integrals)

$f \in RV_\rho$ iff f is locally bounded and $f(tx)/f(x) \to t^\rho$ for every $t > 0$. Equivalently (for the case $\rho < -1$ relevant to Section 3.1): $f \in RV_\rho$ with $\rho < -1$ iff $\int_x^\infty f(t)\,dt$ converges and $\int_x^\infty f(t)\,dt \in RV_{\rho+1}$ with $\int_x^\infty f(t)\,dt \sim xf(x)/|\rho+1|$.

The class Π (de Haan's class). The Gumbel domain (Section 1.4) is sometimes characterized not via von Mises' condition but via the class Π of functions satisfying $[f(x + ta(x)) - f(x)]/f(x) \to \log t$ for some auxiliary function $a(x)$ — a weaker condition than von Mises' differentiability requirement and equivalent to $F \in D(\Lambda)$ under very mild assumptions. This is the formulation used in de Haan and Ferreira's definitive textbook treatment, connecting directly to the auxiliary-function unified criterion of Section 1.4's extension.

Appendix M

Bibliography

Principal references for the theory, methods, and applications developed in Chapter 1 and its appendices, organized thematically. References are cited by author-year throughout the text; this appendix supplies the full bibliographic details.

Extreme Value Theory: Foundations

References

[1] Fisher, R.A. and Tippett, L.H.C. (1928). Limiting forms of the frequency distribution of the largest or smallest member of a sample. *Proceedings of the Cambridge Philosophical Society*, 24, 180–190.

[2] Gnedenko, B.V. (1943). Sur la distribution limite du terme maximum d'une série aléatoire. *Annals of Mathematics*, 44(3), 423–453.

[3] Gumbel, E.J. (1958). *Statistics of Extremes*. Columbia University Press, New York.

[4] Balkema, A.A. and de Haan, L. (1974). Residual life time at great age. *Annals of Probability*, 2(5), 792–804.

[5] Pickands, J. (1975). Statistical inference using extreme order statistics. *Annals of Statistics*, 3(1), 119–131.

[6] de Haan, L. and Ferreira, A. (2006). *Extreme Value Theory: An Introduction*. Springer Series in Operations Research and Financial Engineering. Springer, New York.

[7] Resnick, S.I. (1987). *Extreme Values, Regular Variation, and Point Processes*. Applied Probability, vol. 4. Springer, New York.

[8] Embrechts, P., Klüppelberg, C. and Mikosch, T. (1997). *Modelling Extremal Events for Insurance and Finance*. Applications of Mathematics, vol. 33. Springer, Berlin.

[9] Leadbetter, M.R., Lindgren, G. and Rootzén, H. (1983). *Extremes and Related Properties of Random Sequences and Processes*. Springer Series in Statistics. Springer, New York.

[10] Coles, S.G. (2001). *An Introduction to Statistical Modeling of Extreme Values*. Springer Series in Statistics. Springer, London.

[11] Beirlant, J., Goegebeur, Y., Segers, J. and Teugels, J. (2004). *Statistics of Extremes: Theory and Applications*. Wiley Series in Probability and Statistics. John Wiley & Sons, Chichester.

[12] Gilleland, E. and Katz, R.W. (2016). `extRemes` 2.0: An Extreme Value Analysis Package in R. *Journal of Statistical Software*, 72(8), 1–39.

Regular Variation, Heavy Tails, and Subexponential Distributions

References

[1] Karamata, J. (1930). Sur un mode de croissance régulière des fonctions. *Mathematica (Cluj)*, 4, 38–53.

[2] Bingham, N.H., Goldie, C.M. and Teugels, J.L. (1989). *Regular Variation*. Encyclopedia of Mathematics and its Applications, vol. 27. Cambridge University Press, Cambridge.

[3] Foss, S., Korshunov, D. and Zachary, S. (2013). *An Introduction to Heavy-Tailed and Subexponential Distributions*. 2nd edn. Springer Series in Operations Research and Financial Engineering. Springer, New York.

[4] Nolan, J.P. (2020). *Univariate Stable Distributions: Models for Heavy Tailed Data*. Springer Series in Operations Research and Financial Engineering. Springer, Cham.

[5] Clauset, A., Shalizi, C.R. and Newman, M.E.J. (2009). Power-law distributions in empirical data. *SIAM Review*, 51(4), 661–703.

[6] Meerschaert, M.M. and Sikorskii, A. (2012). *Stochastic Models for Fractional Calculus*. De Gruyter Studies in Mathematics, vol. 43. De Gruyter, Berlin.

Coherent Risk Measures and Financial Risk

References

[1] Artzner, P., Delbaen, F., Eber, J.M. and Heath, D. (1999). Coherent measures of risk. *Mathematical Finance*, 9(3), 203–228.

[2] Rockafellar, R.T. and Uryasev, S. (2000). Optimization of conditional value-at-risk. *Journal of Risk*, 2(3), 21–42.

[3] Föllmer, H. and Schied, A. (2002). Convex measures of risk and trading constraints. *Finance and Stochastics*, 6(4), 429–447.

[4] Kusuoka, S. (2001). On law invariant coherent risk measures. *Advances in Mathematical Economics*, 3, 83–95.

[5] Acerbi, C. (2002). Spectral measures of risk: A coherent representation of subjective risk aversion. *Journal of Banking & Finance*, 26(7), 1505–1518.

[6] Gneiting, T. (2011). Making and evaluating point forecasts. *Journal of the American Statistical Association*, 106(494), 746–762.

[7] Fissler, T. and Ziegel, J.F. (2016). Higher order elicitability and Osband's principle. *Annals of Statistics*, 44(4), 1680–1707.

[8] McNeil, A.J., Frey, R. and Embrechts, P. (2005). *Quantitative Risk Management: Concepts, Techniques and Tools*. Princeton Series in Finance. Princeton University Press, Princeton.

[9] Christoffersen, P.F. (1998). Evaluating interval forecasts. *International Economic Review*, 39(4), 841–862.

Copulas and Dependence Modelling

References

[1] Sklar, A. (1959). Fonctions de répartition à n dimensions et leurs marges. *Publications de l'Institut de Statistique de l'Université de Paris*, 8, 229–231.

[2] Nelsen, R.B. (2006). *An Introduction to Copulas*. 2nd edn. Springer Series in Statistics. Springer, New York.

[3] Joe, H. (1997). *Multivariate Models and Multivariate Dependence Concepts*. Chapman & Hall, London.

[4] Genest, C. and MacKay, R.J. (1986). The joy of copulas: Bivariate distributions with uniform marginals. *American Statistician*, 40(4), 280–283.

[5] Joe, H. (2005). Asymptotic efficiency of the two-stage estimation method for copula-based models. *Journal of Multivariate Analysis*, 94(2), 401–419.

[6] Genest, C. and Rémillard, B. (2008). Validity of the parametric bootstrap for goodness-of-fit testing in semiparametric models. *Annales de l'Institut Henri Poincaré*, 44(6), 1096–1127.

[7] de Haan, L. and Neves, C. (2011). The estimation of the Pickands dependence function using the empirical copula. *Journal of Multivariate Analysis*, 102(6), 1170–1181.

Threshold Methods, Estimation, and Applications

References

[1] Hill, B.M. (1975). A simple general approach to inference about the tail of a distribution. *Annals of Statistics*, 3(5), 1163–1174.

[2] Hosking, J.R.M. and Wallis, J.R. (1987). Parameter and quantile estimation for the generalized Pareto distribution. *Technometrics*, 29(3), 339–349.

[3] Drees, H. and de Haan, L. (1994). Optimal thresholding and the order statistics approach to tail index estimation. Preprint. Full reference in de Haan and Ferreira (2006).

[4] Ferro, C.A.T. and Segers, J. (2003). Inference for clusters of extreme values. *Journal of the Royal Statistical Society, Series B*, 65(2), 545–556.

[5] Martins, E.S. and Stedinger, J.R. (2000). Generalized maximum-likelihood generalized extreme-value quantile estimators for hydrologic data. *Water Resources Research*, 36(3), 737–744.

[6] Davison, A.C. and Smith, R.L. (1990). Models for exceedances over high thresholds (with discussion). *Journal of the Royal Statistical Society, Series B*, 52(3), 393–442.

[7] Coles, S.G. and Tawn, J.A. (1996). A Bayesian analysis of extreme rainfall data. *Applied Statistics*, 45(4), 463–478.

[8] Gomes, M.I., de Haan, L. and Rodrigues, L.H. (2008). Tail index estimation for heavy-tailed models: Accommodation of bias in weighted log-excesses. *Journal of the Royal Statistical Society, Series B*, 70(1), 31–52.

[9] Wang, S. (1998). An actuarial index of the right-tail risk. *North American Actuarial Journal*, 2(2), 88–101.

[10] Yaari, M.E. (1987). The dual theory of choice under risk. *Econometrica*, 55(1), 95–115.

Bayesian Methods in Extreme Value Analysis

References

[1] Stephenson, A. and Tawn, J. (2004). Exploiting occurrence times in likelihood inference for componentwise maxima. *Biometrika*, 92(1), 213–227.

[2] Frigessi, A., Haug, O. and Rue, H. (2002). A dynamic mixture model for unsupervised tail estimation without threshold selection. *Extremes*, 5(3), 219–235.

[3] Padoan, S.A., Ribatet, M. and Sisson, S.A. (2010). Likelihood-based inference for max-stable processes. *Journal of the American Statistical Association*, 105(489), 263–277.

[4] Gelman, A., Carlin, J.B., Stern, H.S., Dunson, D.B., Vehtari, A. and Rubin, D.B. (2014). *Bayesian Data Analysis*. 3rd edn. Chapman & Hall/CRC Press, Boca Raton.

[5] Davison, A.C., Padoan, S.A. and Ribatet, M. (2012). Statistical modeling of spatial extremes (with discussion). *Statistical Science*, 27(2), 161–186.

Non-Stationarity, Climate Extremes, and Operational Risk

References

[1] Coles, S.G., Pericchi, L.R. and Sisson, S. (2003). A fully probabilistic approach to extreme rainfall modeling. *Journal of Hydrology*, 273(1–4), 35–50.

[2] Katz, R.W., Parlange, M.B. and Naveau, P. (2002). Statistics of extremes in hydrology. *Advances in Water Resources*, 25(8–12), 1287–1304.

[3] Milly, P.C.D., Betancourt, J., Falkenmark, M., Hirsch, R.M., Kundzewicz, Z.W., Lettenmaier, D.P. and Stouffer, R.J. (2008). Stationarity is dead: Whither water management? *Science*, 319(5863), 573–574.

[4] Samorodnitsky, G. (2016). *Stochastic Processes and Long Range Dependence*. Springer Series in Operations Research and Financial Engineering. Springer, Cham.

[5] Cruz, M.G., Peters, G.W. and Shevchenko, P.V. (2015). *Fundamental Aspects of Operational Risk and Insurance Analytics*. Wiley Handbooks in Financial Engineering and Econometrics. John Wiley & Sons, Hoboken.

[6] Gutenberg, B. and Richter, C.F. (1956). Earthquake magnitude, intensity, energy, and acceleration. *Bulletin of the Seismological Society of America*, 46(2), 105–145.

Robust Optimization and Decision under Uncertainty (preview of Vol. III)

References

[1] Bertsimas, D. and Sim, M. (2004). The price of robustness. *Operations Research*, 52(1), 35–53.

[2] Delage, E. and Ye, Y. (2010). Distributionally robust optimization under moment uncertainty with application to data-driven problems. *Operations Research*, 58(3), 595–612.

[3] Gao, R. and Kleywegt, A. (2022). Distributionally robust stochastic optimization with Wasserstein distance. *Mathematics of Operations Research*, 48(2), 603–655.

[4] Ben-Tal, A., El Ghaoui, L. and Nemirovski, A. (2009). *Robust Optimization*. Princeton Series in Applied Mathematics. Princeton University Press, Princeton.

Appendix L

Mathematical Prerequisites

A concise treatment of measure theory, integration, characteristic functions, and functional analysis as needed for the proofs and derivations of Chapter 1. Readers fluent in these areas may skim or skip; readers less familiar will find here the specific results used, stated in the forms in which they are applied.

L.1 Probability Spaces and Measurability

Definition (Probability space)

A *probability space* is a triple $(\Omega, \mathcal{F}, P)$ where Ω is a non-empty set (the *sample space*), $\mathcal{F}$ is a σ-algebra on Ω (a collection of subsets closed under complement and countable union, with $\emptyset, \Omega \in \mathcal{F}$), and $P : \mathcal{F} \to [0,1]$ is a probability measure ($P(\Omega) = 1$, P countably additive on disjoint sets).

Definition (Random variable and distribution)

A function $X : \Omega \to \mathbb{R}$ is a *random variable* if it is $\mathcal{F}/\mathcal{B}(\mathbb{R})$-measurable ($\mathcal{B}(\mathbb{R})$ the Borel σ-algebra on $\mathbb{R}$), i.e. $\{X \le x\} \in \mathcal{F}$ for all $x \in \mathbb{R}$. The *distribution* (law) of X is the probability measure μ_X on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ defined by $\mu_X(B) = P(X \in B)$ for $B \in \mathcal{B}(\mathbb{R})$. The *cumulative distribution function* (CDF) is $F_X(x) = P(X \le x) = \mu_X((-\infty, x])$.

CDF properties. Any CDF $F : \mathbb{R} \to [0,1]$ satisfies: (i) F is right-continuous; (ii) $\lim_{x\to-\infty} F(x) = 0$, $\lim_{x\to\infty} F(x) = 1$; (iii) F is non-decreasing. Conversely, any function satisfying (i)–(iii) is the CDF of some probability measure on $\mathbb{R}$ (this is Lebesgue–Stieltjes integration theory).

Quantile function. The *generalized inverse* $F^{-1}(u) = \inf\{x : F(x) \ge u\}$ for $u \in (0,1)$ is the quantile function (Section 1.3). It satisfies: $F^{-1}(u) \le x \iff u \le F(x)$ (for all $u \in (0,1)$ and x at which F is continuous); F^{-1} is left-continuous and non-decreasing; and if $U \sim \text{Uniform}(0,1)$, then $F^{-1}(U) \sim F$ (the probability integral transform, used extensively in copula simulation, Appendix J.2).

L.2 Convergence Concepts

Definition (Convergence modes)

Let $X, X_1, X_2, \ldots$ be random variables on $(\Omega, \mathcal{F}, P)$.

- *Almost sure* (*a.s.*) *convergence*: $X_n \to X$ a.s. if $P(\lim_{n\to\infty} X_n = X) = 1$.
- *Convergence in probability*: $X_n \xrightarrow{P} X$ if $P(|X_n - X| > \epsilon) \to 0$ for all $\epsilon > 0$.
- *Convergence in L^p*: $X_n \xrightarrow{L^p} X$ if $\mathbb{E}[|X_n - X|^p] \to 0$ for $p \ge 1$.
- *Convergence in distribution* (*weak convergence*): $X_n \xrightarrow{d} X$ if $F_{X_n}(x) \to F_X(x)$ at all continuity points x of F_X.

The implications a.s. $\Rightarrow$ in probability $\Rightarrow$ in distribution hold; $L^p \Rightarrow$ in probability. Convergence in distribution does *not* imply convergence in probability in general (it is the weakest mode), yet is the mode appearing throughout Chapters 1–2 (the FTG theorem, the PBdH theorem, the CLT/GCLT of Section 3.4 are all statements of con-

vergence in distribution).

Theorem (Portmanteau theorem for weak convergence)

$X_n \xrightarrow{d} X$ if and only if any one of the following holds: (i) $\mathbb{E}[f(X_n)] \to \mathbb{E}[f(X)]$ for all bounded continuous $f : \mathbb{R} \to \mathbb{R}$; (ii) $\limsup_n P(X_n \in F) \le P(X \in F)$ for all closed sets F; (iii) $\liminf_n P(X_n \in G) \ge P(X \in G)$ for all open sets G; (iv) $P(X_n \in B) \to P(X \in B)$ for all Borel sets B with $P(X \in \partial B) = 0$.

The portmanteau theorem is used implicitly throughout Chapter 1 whenever the FTG limit $F^n(a_n x + b_n) \to G(x)$ is verified at continuity points of G — this is exactly convergence in distribution (condition (iv) with $B = (-\infty, x]$, whose boundary $\partial B = \{x\}$ has $P(X = x) = 0$ whenever G is continuous at x).

Uniform integrability and convergence of moments. If $X_n \xrightarrow{d} X$ and the sequence $\{|X_n|^p\}$ is uniformly integrable (i.e. $\lim_{M\to\infty} \sup_n \mathbb{E}[|X_n|^p \mathbf{1}\{|X_n| > M\}] = 0$), then $\mathbb{E}[|X_n|^p] \to \mathbb{E}[|X|^p]$. This condition is used when going from "convergence in distribution of normalized maxima" to "convergence of moments of normalized maxima" — a non-trivial step requiring the uniform integrability condition on the sequence $M_n/a_n - b_n/a_n$, which fails in general (and explains why, e.g., the mean of M_n properly normalized does not always converge to $\mathbb{E}[G]$, depending on regularity of the tail).

L.3 Characteristic Functions and the Continuity Theorem

Definition (Characteristic function)

The *characteristic function* of a random variable X is $\varphi_X(t) = \mathbb{E}[e^{itX}] = \int e^{itx}\, dF_X(x)$ for $t \in \mathbb{R}$. Properties: $\varphi_X(0) = 1$; $|\varphi_X(t)| \le 1$; φ_X is uniformly continuous. If $\mathbb{E}[|X|^k] < \infty$, then φ_X is k-times differentiable and $\varphi_X^{(k)}(0) = i^k \mathbb{E}[X^k]$.

Theorem (Lévy's continuity theorem)

$X_n \xrightarrow{d} X$ if and only if $\varphi_{X_n}(t) \to \varphi_X(t)$ for all $t \in \mathbb{R}$.

This theorem underlies the proof of the GCLT (Section 3.4): one shows $\varphi_{S_n/a_n}(t) \to \varphi_{X_\alpha}(t)$ (pointwise in t) for the α-stable characteristic function φ_{X_α}, and invokes continuity theorem to conclude $S_n/a_n \xrightarrow{d} X_\alpha$. The same approach works for the CLT ($\alpha = 2$, Gaussian limit) and underlies the "characteristic function method" for proving the FTG theorem as an alternative to the functional-equation proof of Appendix F.1.

Example — Characteristic function approach to the Gumbel limit

For $M_n = \max(X_1, \dots, X_n)$ with $X_i \sim \text{Exp}(1)$, normalizing constants $a_n = 1, b_n = \log n$, so $\tilde{M}_n = (M_n - \log n) \xrightarrow{d} \Lambda$ (Gumbel). The characteristic function of $\tilde{M}_n$:

$$\varphi_{\tilde{M}_n}(t) = \mathbb{E}[e^{it\tilde{M}_n}] = \mathbb{E}[e^{it(M_n - \log n)}] = n^{-it}\mathbb{E}[e^{itM_n}].$$

$\mathbb{E}[e^{itM_n}] = \int_0^\infty e^{itx} n e^{-x}(1-e^{-x})^{n-1} dx$ (using $f_{M_n}(x) = ne^{-x}(1-e^{-x})^{n-1}$ for i.i.d. Exp(1)). Via the substitution $u = 1 - e^{-x}$ and a Laplace-transform type calculation

(standard but lengthy), this evaluates to $\varphi_{\tilde{M}_n}(t) \to \Gamma(1-it)$ as $n \to \infty$, where Γ is the gamma function with complex argument. The characteristic function of the Gumbel distribution Λ is exactly $\varphi_\Lambda(t) = \Gamma(1-it)$ — confirming the convergence, and yielding (as a bonus) the Gumbel's moment formula $\mathbb{E}[\Lambda^k] = \Gamma^{(k)}(1)/(-i)^k$ at $t = 0$, directly reproducing the $\mathbb{E}[X] = \gamma$ (Euler–Mascheroni constant) result via $\Gamma'(1) = -\gamma$.

L.4 Key Results from Real Analysis

The dominated convergence theorem (DCT). If $X_n \to X$ a.s. and $|X_n| \le Y$ a.s. for all n, with $\mathbb{E}[Y] < \infty$, then $\mathbb{E}[X_n] \to \mathbb{E}[X]$. DCT is used throughout Chapter 1 to exchange limits and expectations: e.g. in showing $\mathbb{E}[M_n/a_n] \to \mathbb{E}[G]$ (under uniform integrability), and in the proof of Karamata's theorem (Section 3.1, Appendix F.1).

Monotone convergence theorem (MCT). If $0 \le X_n \nearrow X$ a.s., then $\mathbb{E}[X_n] \nearrow \mathbb{E}[X]$ (no finiteness assumption on $\mathbb{E}[X]$). Used in showing that the GEV mean exists iff $\xi < 1$ (by expressing $\mathbb{E}[X] = \int_0^\infty P(X > x)dx$ and checking convergence via MCT for the integrand).

Fubini–Tonelli theorem. For a product measure space, if $f \ge 0$ (Tonelli) or $\int\int |f|\, d\mu\, d\nu < \infty$ (Fubini), then iterated integrals may be exchanged: $\int\int f\, d\mu\, d\nu = \int\int f\, d\nu\, d\mu$. Used in the proof that $\mathbb{E}[X^k] = \int_0^\infty kx^{k-1}P(X > x)\, dx$ for $X \ge 0$ (by writing $X^k = \int_0^X kx^{k-1}dx$ and exchanging $\mathbb{E}$ with the inner integral via Fubini).

Fatou's lemma. For non-negative X_n: $\mathbb{E}[\liminf X_n] \le \liminf \mathbb{E}[X_n]$. Provides the direction of inequality when DCT does not apply (e.g. showing that $\mathbb{E}[\xi]$ where ξ is a limit of estimators is at most the limit of expected estimates, even when uniform integrability fails).

L.5 Regular Variation: Technical Supplement

This section collects technical results on slowly/regularly varying functions (Section 3.1) used without full proof in the main text.

Theorem (Representation theorem, detailed statement)

$L \in RV_0$ (slowly varying at ∞) if and only if there exist functions $c(x) \to c \in (0, \infty)$ and $\epsilon(x) \to 0$ as $x \to \infty$ such that

$$L(x) = \exp\left(\int_a^x \frac{\epsilon(t)}{t}\, dt\right) \cdot c(x)$$

for some $a > 0$ and all $x \ge a$. In the normalized case (Karamata's canonical form), $c(x) \equiv c$ (constant) and $L(x) = c \exp(\int_a^x \epsilon(t)/t\, dt)$.

Theorem (Uniform convergence on compact sets)

If $L \in RV_0$, then for any $0 < a < b < \infty$:

$$\sup_{t \in [a,b]} \left| \frac{L(tx)}{L(x)} - 1 \right| \to 0 \quad \text{as } x \to \infty.$$

Theorem (Characterization of RV_ρ via integrals)

$f \in RV_\rho$ iff f is locally bounded and $f(tx)/f(x) \to t^\rho$ for every $t > 0$. Equivalently (for the case $\rho < -1$ relevant to Section 3.1): $f \in RV_\rho$ with $\rho < -1$ iff $\int_x^\infty f(t)\,dt$ converges and $\int_x^\infty f(t)\,dt \in RV_{\rho+1}$ with $\int_x^\infty f(t)\,dt \sim xf(x)/|\rho+1|$.

The class Π (de Haan's class). The Gumbel domain (Section 1.4) is sometimes characterized not via von Mises' condition but via the class Π of functions satisfying $[f(x + ta(x)) - f(x)]/f(x) \to \log t$ for some auxiliary function $a(x)$ — a weaker condition than von Mises' differentiability requirement and equivalent to $F \in D(\Lambda)$ under very mild assumptions. This is the formulation used in de Haan and Ferreira's definitive textbook treatment, connecting directly to the auxiliary-function unified criterion of Section 1.4's extension.

Appendix M

Bibliography

Principal references for the theory, methods, and applications developed in Chapter 1 and its appendices, organized thematically. References are cited by author-year throughout the text; this appendix supplies the full bibliographic details.

Extreme Value Theory: Foundations

References

[1] Fisher, R.A. and Tippett, L.H.C. (1928). Limiting forms of the frequency distribution of the largest or smallest member of a sample. *Proceedings of the Cambridge Philosophical Society*, 24, 180–190.

[2] Gnedenko, B.V. (1943). Sur la distribution limite du terme maximum d'une série aléatoire. *Annals of Mathematics*, 44(3), 423–453.

[3] Gumbel, E.J. (1958). *Statistics of Extremes*. Columbia University Press, New York.

[4] Balkema, A.A. and de Haan, L. (1974). Residual life time at great age. *Annals of Probability*, 2(5), 792–804.

[5] Pickands, J. (1975). Statistical inference using extreme order statistics. *Annals of Statistics*, 3(1), 119–131.

[6] de Haan, L. and Ferreira, A. (2006). *Extreme Value Theory: An Introduction*. Springer Series in Operations Research and Financial Engineering. Springer, New York.

[7] Resnick, S.I. (1987). *Extreme Values, Regular Variation, and Point Processes*. Applied Probability, vol. 4. Springer, New York.

[8] Embrechts, P., Klüppelberg, C. and Mikosch, T. (1997). *Modelling Extremal Events for Insurance and Finance*. Applications of Mathematics, vol. 33. Springer, Berlin.

[9] Leadbetter, M.R., Lindgren, G. and Rootzén, H. (1983). *Extremes and Related Properties of Random Sequences and Processes*. Springer Series in Statistics. Springer, New York.

[10] Coles, S.G. (2001). *An Introduction to Statistical Modeling of Extreme Values*. Springer Series in Statistics. Springer, London.

[11] Beirlant, J., Goegebeur, Y., Segers, J. and Teugels, J. (2004). *Statistics of Extremes: Theory and Applications*. Wiley Series in Probability and Statistics. John Wiley & Sons, Chichester.

[12] Gilleland, E. and Katz, R.W. (2016). `extRemes` 2.0: An Extreme Value Analysis Package in R. *Journal of Statistical Software*, 72(8), 1–39.

Regular Variation, Heavy Tails, and Subexponential Distributions

References

[1] Karamata, J. (1930). Sur un mode de croissance régulière des fonctions. *Mathematica (Cluj)*, 4, 38–53.

[2] Bingham, N.H., Goldie, C.M. and Teugels, J.L. (1989). *Regular Variation*. Encyclopedia of Mathematics and its Applications, vol. 27. Cambridge University Press, Cambridge.

[3] Foss, S., Korshunov, D. and Zachary, S. (2013). *An Introduction to Heavy-Tailed and Subexponential Distributions*. 2nd edn. Springer Series in Operations Research and Financial Engineering. Springer, New York.

[4] Nolan, J.P. (2020). *Univariate Stable Distributions: Models for Heavy Tailed Data*. Springer Series in Operations Research and Financial Engineering. Springer, Cham.

[5] Clauset, A., Shalizi, C.R. and Newman, M.E.J. (2009). Power-law distributions in empirical data. *SIAM Review*, 51(4), 661–703.

[6] Meerschaert, M.M. and Sikorskii, A. (2012). *Stochastic Models for Fractional Calculus*. De Gruyter Studies in Mathematics, vol. 43. De Gruyter, Berlin.

Coherent Risk Measures and Financial Risk

References

[1] Artzner, P., Delbaen, F., Eber, J.M. and Heath, D. (1999). Coherent measures of risk. *Mathematical Finance*, 9(3), 203–228.

[2] Rockafellar, R.T. and Uryasev, S. (2000). Optimization of conditional value-at-risk. *Journal of Risk*, 2(3), 21–42.

[3] Föllmer, H. and Schied, A. (2002). Convex measures of risk and trading constraints. *Finance and Stochastics*, 6(4), 429–447.

[4] Kusuoka, S. (2001). On law invariant coherent risk measures. *Advances in Mathematical Economics*, 3, 83–95.

[5] Acerbi, C. (2002). Spectral measures of risk: A coherent representation of subjective risk aversion. *Journal of Banking & Finance*, 26(7), 1505–1518.

[6] Gneiting, T. (2011). Making and evaluating point forecasts. *Journal of the American Statistical Association*, 106(494), 746–762.

[7] Fissler, T. and Ziegel, J.F. (2016). Higher order elicitability and Osband's principle. *Annals of Statistics*, 44(4), 1680–1707.

[8] McNeil, A.J., Frey, R. and Embrechts, P. (2005). *Quantitative Risk Management: Concepts, Techniques and Tools*. Princeton Series in Finance. Princeton University Press, Princeton.

[9] Christoffersen, P.F. (1998). Evaluating interval forecasts. *International Economic Review*, 39(4), 841–862.

Copulas and Dependence Modelling

References

[1] Sklar, A. (1959). Fonctions de répartition à n dimensions et leurs marges. *Publications de l'Institut de Statistique de l'Université de Paris*, 8, 229–231.

[2] Nelsen, R.B. (2006). *An Introduction to Copulas*. 2nd edn. Springer Series in Statistics. Springer, New York.

[3] Joe, H. (1997). *Multivariate Models and Multivariate Dependence Concepts*. Chapman & Hall, London.

[4] Genest, C. and MacKay, R.J. (1986). The joy of copulas: Bivariate distributions with uniform marginals. *American Statistician*, 40(4), 280–283.

[5] Joe, H. (2005). Asymptotic efficiency of the two-stage estimation method for copula-based models. *Journal of Multivariate Analysis*, 94(2), 401–419.

[6] Genest, C. and Rémillard, B. (2008). Validity of the parametric bootstrap for goodness-of-fit testing in semiparametric models. *Annales de l'Institut Henri Poincaré*, 44(6), 1096–1127.

[7] de Haan, L. and Neves, C. (2011). The estimation of the Pickands dependence function using the empirical copula. *Journal of Multivariate Analysis*, 102(6), 1170–1181.

Threshold Methods, Estimation, and Applications

References

[1] Hill, B.M. (1975). A simple general approach to inference about the tail of a distribution. *Annals of Statistics*, 3(5), 1163–1174.

[2] Hosking, J.R.M. and Wallis, J.R. (1987). Parameter and quantile estimation for the generalized Pareto distribution. *Technometrics*, 29(3), 339–349.

[3] Drees, H. and de Haan, L. (1994). Optimal thresholding and the order statistics approach to tail index estimation. Preprint. Full reference in de Haan and Ferreira (2006).

[4] Ferro, C.A.T. and Segers, J. (2003). Inference for clusters of extreme values. *Journal of the Royal Statistical Society, Series B*, 65(2), 545–556.

[5] Martins, E.S. and Stedinger, J.R. (2000). Generalized maximum-likelihood generalized extreme-value quantile estimators for hydrologic data. *Water Resources Research*, 36(3), 737–744.

[6] Davison, A.C. and Smith, R.L. (1990). Models for exceedances over high thresholds (with discussion). *Journal of the Royal Statistical Society, Series B*, 52(3), 393–442.

[7] Coles, S.G. and Tawn, J.A. (1996). A Bayesian analysis of extreme rainfall data. *Applied Statistics*, 45(4), 463–478.

[8] Gomes, M.I., de Haan, L. and Rodrigues, L.H. (2008). Tail index estimation for heavy-tailed models: Accommodation of bias in weighted log-excesses. *Journal of the Royal Statistical Society, Series B*, 70(1), 31–52.

[9] Wang, S. (1998). An actuarial index of the right-tail risk. *North American Actuarial Journal*, 2(2), 88–101.

[10] Yaari, M.E. (1987). The dual theory of choice under risk. *Econometrica*, 55(1), 95–115.

Bayesian Methods in Extreme Value Analysis

References

[1] Stephenson, A. and Tawn, J. (2004). Exploiting occurrence times in likelihood inference for componentwise maxima. *Biometrika*, 92(1), 213–227.

[2] Frigessi, A., Haug, O. and Rue, H. (2002). A dynamic mixture model for unsupervised tail estimation without threshold selection. *Extremes*, 5(3), 219–235.

[3] Padoan, S.A., Ribatet, M. and Sisson, S.A. (2010). Likelihood-based inference for max-stable processes. *Journal of the American Statistical Association*, 105(489), 263–277.

[4] Gelman, A., Carlin, J.B., Stern, H.S., Dunson, D.B., Vehtari, A. and Rubin, D.B. (2014). *Bayesian Data Analysis*. 3rd edn. Chapman & Hall/CRC Press, Boca Raton.

[5] Davison, A.C., Padoan, S.A. and Ribatet, M. (2012). Statistical modeling of spatial extremes (with discussion). *Statistical Science*, 27(2), 161–186.

Non-Stationarity, Climate Extremes, and Operational Risk

References

[1] Coles, S.G., Pericchi, L.R. and Sisson, S. (2003). A fully probabilistic approach to extreme rainfall modeling. *Journal of Hydrology*, 273(1–4), 35–50.

[2] Katz, R.W., Parlange, M.B. and Naveau, P. (2002). Statistics of extremes in hydrology. *Advances in Water Resources*, 25(8–12), 1287–1304.

[3] Milly, P.C.D., Betancourt, J., Falkenmark, M., Hirsch, R.M., Kundzewicz, Z.W., Lettenmaier, D.P. and Stouffer, R.J. (2008). Stationarity is dead: Whither water management? *Science*, 319(5863), 573–574.

[4] Samorodnitsky, G. (2016). *Stochastic Processes and Long Range Dependence*. Springer Series in Operations Research and Financial Engineering. Springer, Cham.

[5] Cruz, M.G., Peters, G.W. and Shevchenko, P.V. (2015). *Fundamental Aspects of Operational Risk and Insurance Analytics*. Wiley Handbooks in Financial Engineering and Econometrics. John Wiley & Sons, Hoboken.

[6] Gutenberg, B. and Richter, C.F. (1956). Earthquake magnitude, intensity, energy, and acceleration. *Bulletin of the Seismological Society of America*, 46(2), 105–145.

Robust Optimization and Decision under Uncertainty (preview of Vol. III)

References

[1] Bertsimas, D. and Sim, M. (2004). The price of robustness. *Operations Research*, 52(1), 35–53.

[2] Delage, E. and Ye, Y. (2010). Distributionally robust optimization under moment uncertainty with application to data-driven problems. *Operations Research*, 58(3), 595–612.

[3] Gao, R. and Kleywegt, A. (2022). Distributionally robust stochastic optimization with Wasserstein distance. *Mathematics of Operations Research*, 48(2), 603–655.

[4] Ben-Tal, A., El Ghaoui, L. and Nemirovski, A. (2009). *Robust Optimization*. Princeton Series in Applied Mathematics. Princeton University Press, Princeton.